Analysis of High Frequency Smart Meter Energy Consumption Data

Tureczek, Alexander Martin

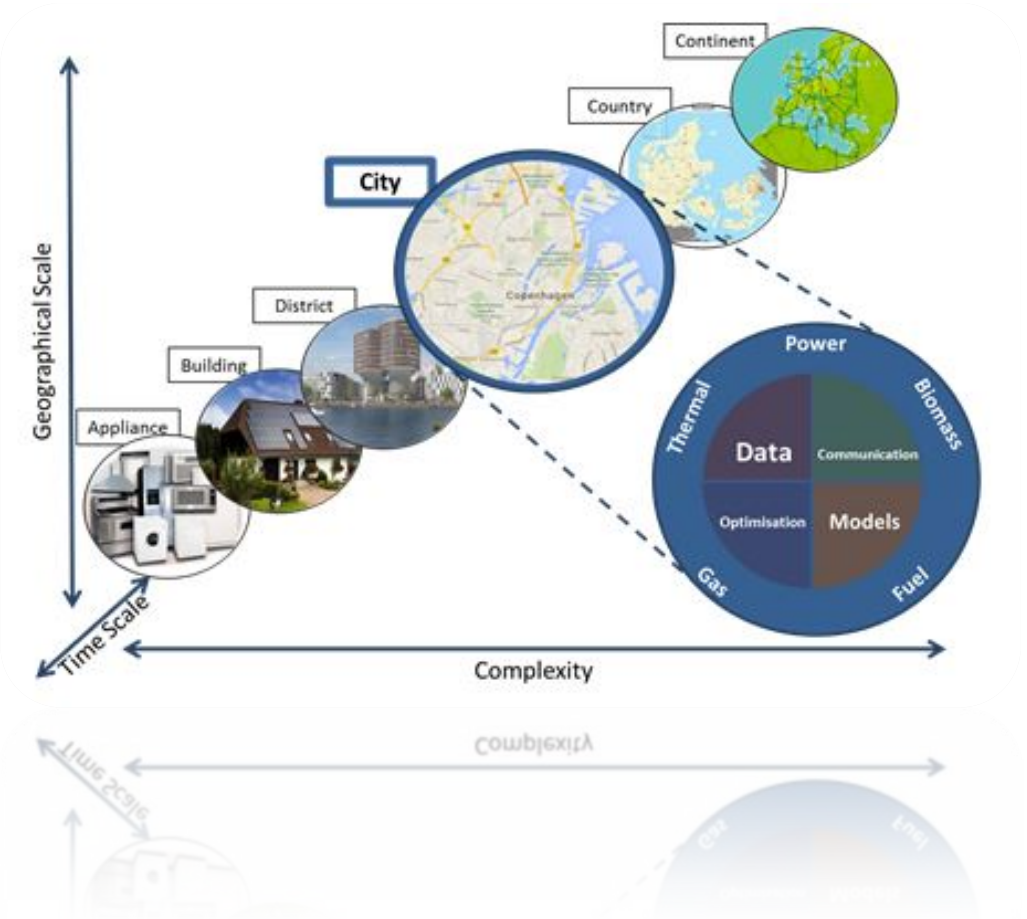
Publication date:
2019

Document Version
Publisher's PDF, also known as Version of record

Ph.D. Thesis

Analysis of High Frequency Smart Meter Energy Consumption Data

Alexander Martin Tureczek



Technical University of Denmark
Department of Management Engineering, Systems Analysis
Produktionstorvet, Building 426, room 125.
2800 Kgs Lyngby, Denmark.
July 2018

Supervisors:
Per Sieverts Nielsen, DTU Management Engineering and Henrik Madsen, DTU Compute

Page intentionally left blank.

Not everything that can be counted counts,

and not everything that counts can be counted

William Bruce Cameron, 1963

Preface

This Ph.D. project entitled “Analysis of High-Frequency Smart Meter Energy Consumption Data” has been conducted at the Department of Management Engineering at the Technical University of Denmark (DTU) from August 2015 to July 2018. The thesis has been submitted to the Department of Management Engineering in partial fulfillment of the requirements of acquiring the Ph.D. degree. The work is co-founded by DTU and the Danish Innovation Found as part of the interdisciplinary research project CITIES and was supervised by Senior Researcher Per Sieverts Nielsen (DTU) and Professor Henrik Madsen (DTU).

The thesis investigates the applicability of smart meter electricity data and smart meter district heating data for consumption clustering, it is divided into two parts: part one introduces the thesis background and motivation. It includes a brief overview of the theory applied, the results obtained and a discussion of the results in a larger setting. The latter part is a collection of the four research papers that have been written during the Ph.D. study.

Alexander Martin Tureczek, July 2018.

Acknowledgement

This Ph.D. study has been made possible through the CITIES project, which involves many people a lot of whom I've never met, but without everyone's commitment and drive I would never have had this opportunity. Thank you so much and I hope this dissertation meets the expectations!

I would like to express my gratitude to my supervisor Senior Researcher Per Sievert Nielsen and co-supervisor Professor Henrik Madsen for guidance, feedback and valuable input throughout this project, and especially for this opportunity and for believing in me. And thank you Per for all the non-scientific discussions about betting, sailing, running etc. it really helped me keeping things in perspective.

I wish to thank SydEnergi for granting me access to such a large and nicely structured dataset. I want to express my deepest gratitude to the people at AffaldVarme Aarhus for opening their doors and letting a complete stranger be part of their inspiring environment. And a special thanks to Adam Brun for valuable advice and showing me how and what is important in district heating and Magnus Dahl and Rasmus Pedersen for answering all my questions about district heating and thermodynamics. It really helped!

Thank you to Niels Framroze Møller and the rest of the QED discussion group, to Frits Møller Andersen, Ida Græsted Jensen and Lise Skovsgaard Nielsen for always taking the time for a chat when I was roaming the halls in building 426 frustrated by some problem or just in need of some advice. It always helped clear my mind and I enjoyed every moment and conversation. Thank you, Rikke Brinkø Berg, for sharing the secrets of Ph.D. survival with me; your advice helped me overcome the frustrations of Ph.D. life. Thank you Per Skougaard Kaspersen and Morten Andreas Dahl Larsen for all the nice runs around campus (including CPH-Half), fantasy manager and Thursday morning football. To Xiufeng Liu, Kirsten Halsnæs and the rest of the colleagues at Systems Analysis, CCSD, and in the CITIES project, thank you for helping me understand the energy system and the many good discussions, valuable inputs and occasional Friday beer you made this journey a little less lonely and much more fun.

And to my family and friends thank you for your support, encouragement, and helping me find motivation along the way. I would like to thank my wife Pia Thomsen for encouraging me and letting me reach out for this opportunity, thank you for all our discussions, your input and endless evenings of proofreading.

This work was in part supported by the Danish Innovation Found under Grant DSF 1305-00027B.

English Summary

As society moves towards increasing electrification in areas such as transportation, the future peak electricity demand may very well exceed the capacity of the electricity grid. Consumption flexibility is expected to play an important role in peak shawing and smart meters can help analyze demand. Electricity smart meters are capable of recording consumption at very high frequency, down to the minute. These recordings allow for unprecedented consumption insights and identification of consumption patterns and flexibility. This thesis investigates the ability of electricity smart-meter consumption data to be used for consumption clustering to identify consumer types and enable diverse tariff structures and thus incentivize flexible consumption patterns.

Through a systematic literature review the state of the art in smart meter consumption clustering is outlined and evaluated, the systematics of the review ensure reproducibility of the results. The review identifies that simple methods such as K-Means and Hierarchical clustering are prevailing; though more advanced methods are applied but their complexity and lack of improved cluster structures render them as unpopular choices. The review recognizes that smart-meter consumption data collected for billing purposes are applicable for clustering, but that the clusters are ambiguous, and their long-term stability is questionable.

The lessons from the review are applied to a Danish electricity consumption dataset containing readings from more than 32,000 smart meters. The results obtained from the Danish data are comparable to international studies of electricity smart-meter consumption data. Furthermore, the analysis of the data introduces autocorrelation features to successfully improve the clustering potential of K-Means to include temporal dependencies. The clusters produced are still ambiguous but clustering is finer grained and within-cluster variance is reduced. It is investigated if the results from the review and the electricity data are readily applicable for clustering of smart-meter district heating consumption data. The methods used for electricity data are successfully applied to cluster consumption for district heating heat exchange stations, without change in methodology. The results are similar to those of electricity consumption clustering with equivalent conclusions regarding clustering of consumption data with temporal components.

This thesis further investigates the time stability of the developed clusters by introduction of a novel methodology; Varatio able to evaluate if households are clustered together over time. Varatio applies variance ratios to compare clustering solutions. The analysis of cluster stability shows that the smart meter consumption clusters produced by K-Means are highly unstable, with stability of clusters being less than 20% of the meters.

The thesis concludes that smart-meter data can be applied to identify consumption clusters, but the current prevailing methodology produce academically viable clusters with limited practical applicability. There are structures in the data that the methodology currently applied are unable to manage e.g. reduce the within cluster variance to such a degree that the clusters are uniquely defined and identifiable. Further research into methods for time series clustering is needed to control the cluster variance and enable distinct consumption clusters.

Danish Summary

Den stigende elektrificering af samfundet i særdeleshed transport, vil have stor indflydelse på fremtidens spidsbelastning af el nettet, og vil i perioder kunne overstige nettets kapacitet. Det er forventet at forbrugsfleksibilitet vil komme til at spille en vigtig rolle i reduktion af spidsbelastningen. Smarte elmålere, smart meters, er i stand til at aflæse forbrug på minut basis. Disse målere åbner for forbrugsmålinger på et hidtil uset detaljeniveau til brug for identificering af forbrugsmønstre og fleksibilitet. Denne afhandling undersøger mulighederne for at anvende elforbrugsmålinger til kategorisering af forbrug med deraf følgende forbrugstyper. Dette kan assistere i udvikling af specifikke el-abonnementer som kan tilskynde til forbrugsfleksibilitet.

Gennem en systematisk litteratur gennemgang evalueres forskningsresulter i energi forbrugskategorisering. Systematikken sikre at resultaterne er reproducerbare. Litteratur gennemgangen identificerer K-Means og Hierarkisk clustering som de mest anvendte metoder til kategorisering af energi forbrug. Mere avancerede metoder er anvendes sporadisk, men deres kompleksitet opvejer ikke de marginale forbedringer i kategorierne. Gennemgangen finder også at smart meter forbrugsdata kan anvendes til at identificere forbrugsmønstre, men at de identificerede mønstres stabilitet over tid er tvivlsom.

Læring fra litteraturgennemgangen er anvendt til analyse af Danske elforbrugsdata fra mere end 32.000 husstande. Resultaterne af analyserne af danske data er identiske med resultaterne fra sammenlignelige internationale studier. Dertil har analyserne udført i denne afhandling introduceret autokorrelation features til at forbedre K-Means clusteringen ved at inkludere autokorrelation. De kategorier som bliver identificeret i de danske data er ikke unikke, grundet kategori-variation som resulterer i overlap mellem grupperne. Det undersøges også om metoderne fra elforbrugsanalyserne direkte kan anvendes på fjernevarmeforbrugsdata. Det konkluderes at fjernvarmeforbrugsdata kan forbrugsklassificeres med samme metoder som anvendes til elforbrugsklassificering. Resultaterne for fjernevarmedata er identiske med resultaterne for elforbrugsdata, og konklusionerne for autokorrelation ligeså.

Ydermere evaluerer denne afhandling stabiliteten af de identificerede forbrugskategorier via en nyudviklet metode; Varatio, som er i stand til at analyserer om kategorierne er stabile over tid. Varatio anvender varians forhold til at sammenligne forbrugskategorier over tid. Analysen af stabiliteten viser at forbrugskategorierne beregnet med K-Means er ustabile over tid.

Denne afhandling konkluderer at el- og fjernvarmeforbrugsdata fra digitale smart-målere kan anvendes til at identificere forbrugskategorier. Men at den for tiden fremherskende metode kan beregne forbrugskategorier så er den praktiske anvendelse begrænset. Der er for stor variation i de enkelte grupper hvilket resulterer i at grupperne overlapper. Dette skyldes at der er underliggende strukturer i data som de anvendte metoder ikke er i stand til at håndtere. For at kunne generere unikke forbrugskategorier er mere forskning i tidsrække klassificering nødvendigt.

Table of Contents

Figures

Tables

List of publications

Articles included in the thesis.

Paper 1 - Tureczek. Alexander M, Nielsen. Per S. Structured Literature Review of Electricity Consumption Classification Using Smart Meter Data. Energies 2017;10:584. doi:10.3390/en10050584.

Paper 2 - Tureczek A, Nielsen PS, Madsen H. Electricity Consumption Clustering Using Smart Meter Data. Energies 18AD;11:859. doi:10.3390/en11040859.

Paper 3 - Alexander Martin Tureczek, Per Sieverts Nielsen, Henrik Madsen, Adam Brun. Clustering District Heat Exchange Stations Using Smart Meter Consumption Data. Published in Energy & Buildings (October 2018).

Paper 4 – Tureczek. Alexander M,. Stability of Electricity Smart Meter Consumption Clusters over Time. Submitted to Elsevier Applied Energy (July 2018).

Essential Terms and Definitions

Alphabetically ordered

ACF – **Autocorrelation Function** – a measure of time dependence in data. In smart meter consumption data, it measures the influence of previous consumption reading to the current. It does so for a specified time step, e.g. hourly and for a specified period e.g. 24 hours.

AVA – **AffaldVarme Aarhus** – At municipality owned district heating supplier. AVA is supplying greater Aarhus with district heating. As partner in the CITITES project AVA has kindly provided this thesis with consumption analyzed in paper 3.

CDI – **Cluster Dispersion Index** – A method for evaluating similarities between clusters created by for instance K-Means. See CVI.

CITIES – **Center for IT-Intelligent Energy Systems in Cities** - The research this thesis is part of. The overall aim of the project is an integrated analysis of the Danish energy system encompassing all aspects from production to consumption.

CVI – **Cluster Validation Index** – a general term for a family of metrics evaluating cluster solutions from the K-Means algorithm.

Clustering - the statistical process of subdividing data into homogeneous groups which exhibit smaller variance than the original data. Clusters are mathematically distinguishable via clustering methods. Each cluster in a clustering solution has cluster profile or cluster definition.

Cluster Interval – In clustering often the optimal number of clusters is unknown. Therefor an interval in the range of expected optimal is traversed to identify the optimal cluster number. In this thesis this interval is denoted K and the optimal is denoted k*.

Data cleaning – **data cleansing** – is the process of refining the data to achieve analytical quality. This includes identification of missing, outliers and the handling of such cases.

DBI – **Davies-Bouldin Index** - A method for evaluating similarities between clusters created by for instance K-Means. See CVI.

DC - **District Heating** – An efficient heating technology using water a medium for distributing heating to large areas of buildings.

DSO – **Distribution System Operator** – A utility responsible for the electricity or district heating distribution network. The part of the energy system that reaches end-users.

GDPR – **General Data Protection Regulation** – A European Union legislation on data protection and privacy, which utilities must adhere to.

HX – **Heat Exchange Station** – The district heating exchange stations are large substations converting 120^{o}C transmission grid water to 80^{o}C distribution grid water for the end-users. It is the district heating equivalent to electricity transformation stations.

MIA – **Mean Index Adequacy** - A method for evaluating similarities between clusters created by for instance K-Means. See CVI.

Preprocessing of data – is the process of preparing data for analysis. In this thesis preprocessing is a data transformation which extracts features from the data prior to clustering. In contrast to "data cleaning" which refines to achieve analytical quality. See data cleaning.

SE – **SydEnergi** – SE is an electricity utility supplying large parts of southern Denmark with electricity. As partner in the CITITES project SE has kindly provided this thesis with consumption analyzed in paper 2 and 4.

TSO – **Transmission System Operator** – The operator of the high-voltage electricity grid, or the 120^{o}C District heating water in Denmark.

Page intentionally left blank.

PART I - SMART METER DATA ANALYSIS

Page intentionally left blank.

1. Introduction

This thesis is divided into seven independent chapters each is self-containing and contributing to the overall results and conclusion. Chapter 1 outlines the motivation of the research including a brief description of the long-term targets of the Danish energy system, and consumption clustering. Furthermore, this chapter presents the research questions of the thesis, the four papers included and the contributions that the Ph.D. study conveys. Chapter 2 introduces the concept of smart metering, describes the datasets applied throughout the papers and outlines the software used to achieve the results. The theoretical framework applied is outlined in chapter 3 including novel contributions of the thesis, especially in section 3.4 which presents Pseudo Cross-Validation and section 3.7 presenting Varatio. Chapter 4 outlines the results of the four papers along with a brief description of each papers scientific outline, methodology, results and conclusions. The results are discussed in chapter 5, followed by a conclusion and outlook in chapter 6.

1.1. Motivation and Background

Denmark has set itself a long-term target of phasing out its use of fossil fuels by introducing 100% renewable energy by 2050 [1]. Many different technologies are earmarked as playing an integral part in the Danish transition. The current Danish energy mix consists of several different renewable and fossil-based sources, with substantial penetrations of wind energy [2], biomass [3] and solar fields [4][5]. The large amount of wind energy introduces high volatility into electricity-generating capacity. Electricity generation is expected to be 84% reliant on renewable sources by 2020 [2], which even now periodically exceed demand [6].

Without technology to store electricity large volatility in electricity production requires high degree of flexibility in consumption as electricity must be consumed when produced. Flexibility in consumption can reduce maintenance costs, limit grid expansion and overall better use of resources, especially with volatile electricity production. It is believed that consumption flexibility can be motivated through differences in individual household consumption profiles and that smart meters may hold the key to identifying consumption archetypes. Furthermore, increased electrification of the future society also increases the need for flexible consumption to avoid extensive and expensive grid investments to cope with additional consumption [7].

The Danish energy-supply system is divided into two major components, electricity and district heating (DH). District heating systems are extensively developed in all major cities and many rural districts, with a total of 430 district heating utilities and more than 60,000 kilometers of pipes. The total penetration of DH in Denmark is 64% of households [8], and incorporates large pit-storage facilities which ensure stable heat supply throughout the year [9], [10]. The large penetration of DH, making the Danish energy system globally unique, significantly reduces the overall need for electricity consuming heat pumps. The reduced demand for heat pumps also reduces household flexibility as heating is a significant part of household energy consumption. Furthermore the lack of heat pumps in Denmark also impacts the total electricity consumption in comparison to countries using electricity for heating.

Previous research projects have focused on separate aspects of the energy system, consequently neglecting the overall effects of cost, consumption, and production optimization that are achievable through an integrated approach to energy flexibility analysis, encompassing the entire system. Notable research projects focusing on the individual aspects include IPower [11], Ensymora [12], EcogridEU [13], Flexpower [14] and 4DH [15]. All but 4DH focus on electricity, while 4DH focus on district heating.

The **C**enter for **IT**-**I**ntelligent **E**nergy **S**ystems in Cities (CITIES) research project [16], of which this thesis forms a part, aims to establish an integrated approach to energy systems analysis in order to harness the flexibility of combined energy-efficiency analyses, rather than optimizing individual aspects of the energy mix. The CITIES project consists of seven work packages (WPs), each with their own specific focus: 1) Demand, 2) Production, 3) Integration, 4) Aggregation and Markets, 5) Forecasting and Control, 6) Simulation and Planning, and 7) Decision Support. These seven WPs evaluate the entire energy system from production to distribution and final consumption. This enables the energy system to be analyzed as a single integrated entity and the flexibility gained through an integrated approach to be assessed. This thesis is embedded in WP1 – Demand, which through a data driven approach, contributes to the subtask of analyzing consumption characteristics and profiles for evaluating end-user consumption flexibility.

Digital technologies are predicted to be an essential component in understanding and optimizing the energy system and consumption behavior. An important technology that the modern energy system uses is smart meters. One prospective task of smart-metering is to unveil consumption patterns in order to facilitate the identification of end-user flexibility. Danish electricity utilities are required by law to install smart meters with all consumers by the end of 2020 [17]. These meters are capable of monitoring consumption at very high frequencies, down to the minute. The resulting data allow household consumption to be monitored in unprecedented detail in order to integrate the end-user much more closely to the energy system, not only as a consumer, but also as a supplier of information on consumption and demand. The widespread installation of smart meters across the European Union is expected to exceed 72% in 2020 and to initiate energy savings immediately [18], [19].

Although, with the European Union's general data protection regulation (GDPR), smart-meter data increases operational complexity for the utility companies, smart meters are expected to play a vital role in how utilities operate and optimize. Especially when identifying consumption flexibility and for the purpose of tariff development. Currently the smart-meter data that are collected and stored are used for the automatic billing of consumption. Few utilities use the data for consumer engagement, for example, by enabling households to view their consumption through mobile phone apps [20]. Society and utilities at large are interested in identifying and assessing the potential of smart-meter data beyond automatic billing. The promised potential from smart meters for demand flexibility or customer acquisition, whether through value propositions such as apps, tailored tariff structures catering for specific customers or the ability to have optimal control of appliances, has yet to be realized [21]–[24].

The application of smart-meter consumption data for analysis is a relatively new research field, predominantly addressing household electricity consumption behavior in countries that use electricity for heating. A few regions around the globe have been driving the research. In Europe, Italy has been driving the initial research into smart-meter data acquisition through the early installation of smart meters. In particular, [25] and [26] have been extensively cited for their initial analyses of smart-meter data for consumption profiling to aid in tariff development. The Republic of Ireland has conducted a study of approximately 4,000 dwellings, which have been combined with survey data, and subsequently opening this material to researchers [27] and [28]. Japan and South Korea have likewise combined survey data with smart-meter data for profiling consumer types within clusters [29]–[31]. Also, studies of consumer profiling have been undertaken in the United States and Canada [32]–[34]. The common feature in the assessed papers is the use of electricity for heating. The papers all successfully apply smart-meter data to cluster consumers for investigation of consumption flexibility, although the resulting clusters may be academically viable their practical applicability is questionable [35]. The applicability of the results in a Danish setting with the large penetration of district heating is unexplored and therefore not necessarily evident.

This thesis contributes to the literature of smart meter consumption clustering by applying the state of the art in consumption clustering to the two major components of the Danish energy system. It does so by showing the applicability of the methods for electricity and district heating consumption data from the Danish energy system. Through a novel framework, pseudo cross-validation, I extend cross-validation to the cluster selection process of K-Means thereby reducing selection bias for increased solution generalizability. Furthermore, I present Varatio, a novel method for easy evaluation of the persistence of clusters across time. The methods aid in generalizability and evaluation of stability of the consumption cluster solutions.

1.2. Scope of the thesis

The CITIES research project aims at identifying flexibility in the energy system from production to end-use. As part of WP1 the scope of this dissertation was originally envisioned to analyze consumer behavior, and identify consumption flexibility using smart meter energy consumption data from Danish utilities. Paper 1 identified nine different themes in the field of smart meter consumption data analysis shown in Figure 1. The borderline group contains papers where the preliminary analysis could not establish if smart meter data were used in the study.

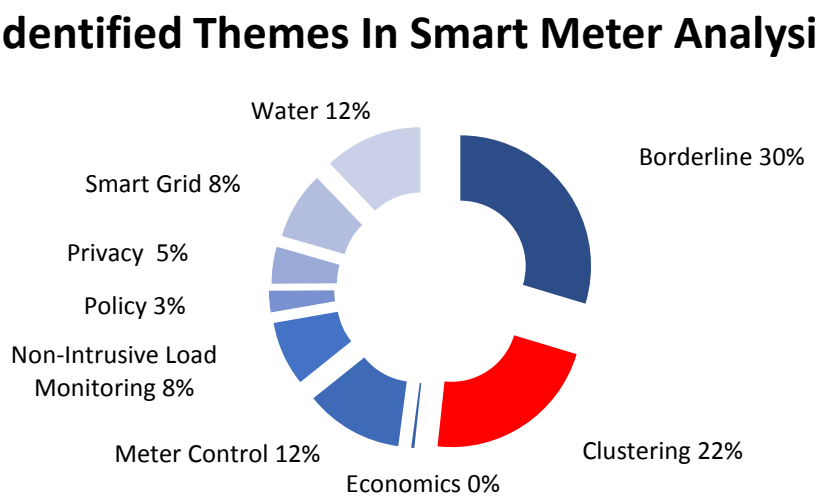


Figure 1 –– Smart meter data analysis themes identified in paper 1. Clustering of consumption is the main focus of 22% of the studies analyzed. It will also be the focus of this PhD study.

The analysis of end-user consumption at household level is a tremendous task. A simpler task is to identify households which behave in similar fashion, enabling development of consumer profiles which encompass groups of consumers allowing for the analysis of consumers types rather than individuals. Though the title of this thesis could encompass each of the nine topics, clustering (red) was selected as the most relevant for end-user behavior and flexibility analysis as part of WP1. Therefore, clustering consumers together in homogeneous consumer types has been the main focus of the thesis.

During the process of identifying and assessing relevant literature in paper 1, it became apparent that the methods employed in the current state of the art in profiling energy consumption, using smart meter data, are inadequate. The clustering of consumers proved to be a non-trivial task requiring methodology capable of exploiting intrinsic information in data, methodology which does not yet exist. This insight has shifted the focus of this project from solely analyzing energy consumer behavior, to also identifying and alleviating

gaps in the methodology for energy consumption clustering, with subsequent application to the Danish case in papers 2, 3 and 4.

This dissertation evaluates end-user behavior in the Danish energy system by using high-frequency smart-meter energy consumption data to evaluate consumer behavior. Consumption behavior has been analyzed through the acquisition of smart-meter electricity and district-heating data from SydEnergi (SE) and AffaldVarme Aarhus (AVA) respectively. The composition of the Danish energy system, with its reduced electricity demand for heating outlined above, presents a different scenario for consumption clustering than examined in other studies. This makes the thesis and its contributions novel and relevant to the current literature on consumption clustering.

This thesis will focus solely on consumption clustering. The selected scope is unfortunately limiting the opportunity for a thorough and comprehensive discussion of several important topics related to the application of smart meter consumption data. Topics such as; forecasting, identification of specific appliances (NILM), legislation and data privacy, are out of scope. The thesis does however fully acknowledge the importance of each of these topics. Even within the scope, this dissertation cannot produce an exhaustive analysis of smart meter data's capability to identify consumption archetypes. It is a contribution to the field of smart meter consumption clustering and clustering methodology in general, with offset in K-Means clustering.

1.3. Research Objectives

This thesis revolves around five research objectives regarding smart-meter data for consumption clustering. The objectives form a natural progression in that they examine successively the applicability of smart-meter consumption data (Q1) current research (Q2), the application of existing knowledge to Danish smart meter data (Q3, Q4), and finally the development of novel methods that drive the field forward (Q5).

Q 1) Is it possible to cluster electricity consumption patterns using smart-meter electricity or district heating time-series data collected for billing purposes? This question is focused on identifying the potential for utilities to apply smart-meter consumption data for clustering. A general question underlining papers 1, 2, 3, and 4.

Q 2) Analysis and evaluation of the current state of the art in smart-meter consumption clustering. This objective covers the identification of the prevailing methodology, prominent data sets and pioneering papers in this field, as well as identification of potential gaps and pitfalls. Paper 1 conducts a systematic literature review, analyzing and critically evaluating the field.

Q 3) Application of the prevailing methodology identified in (Q2) applied to Danish smart-meter electricity data for consumption clustering. Paper 2 analyzes such data, applies the main methods uncovered in (Q2) and attempts to alleviate the identified pitfalls.

Q 4) Application of the prevailing methodology identified in (Q2) applied to district-heating smart-meter data for consumption clustering in Denmark. Paper 3 analyzes data from district-heating exchange stations to evaluate the applicability of the methodology from electricity smart-meter clustering in a district-heating setting. Suggestions are made for alleviating and circumventing the gaps and pitfalls identified for smart-meter clustering in (Q2).

Q 5) Evaluation of the stability of attained clustering solutions for electricity data. This objective concerns the applicability of consumption clustering and the stability of the solution for electric utilities looking at smart-meter clustering for value propositions. Paper 4 evaluates cluster stability over time and develops methodologies for estimating it.

The ambition of the five research objectives is to investigate the potential application of smart-meter data for consumption clustering. Investigation of the current literature on smart meter consumption clustering enables identification of gaps and potential pitfalls. Applying the knowledge from the literature to Danish consumption data will confirm if it is applicable in a Danish setting. As district heating is widespread in Denmark, the applicability to heating profiles is relevant. Finally the thesis will investigate how to assess the stability of achieved cluster solutions; this is needed for generalizability and to promote the applicability of the cluster solutions outside academia.

1.4. Peer-Reviewed Journal Papers Submitted

The foundation of this thesis is the four research papers that are submitted and appended to this dissertation. Papers 1, 2, and 3 have already been published, and paper four is being submitted for review at the present time. Below is a brief description of the main objectives of each paper.

Paper 1) **Structured Literature Review of Electricity Consumption Classification Using Smart Meter Data (published)**. On an adaptation of Okoli's methodology for systematic review and subsequent investigation of the current trends and prevailing methodologies in consumption clustering. The systematic review process conducted here ensures the reproducibility of the paper findings and evaluation.

Paper 2) **Electricity Consumption Clustering Using Smart Meter Data (published)**. The methodologies identified in paper 1 are evaluated in relation to Danish smart-meter electricity data. Efforts are directed at reproducing the results within a Danish setting and filling the gaps identified in paper 1. This paper applies prominent methods from the literature and introduces novel applications of statistical and mathematical methods to improve consumption clustering.

Paper 3) **Clustering District Heat Exchange Stations Using Smart Meter Consumption Data (published)**. Using the insights gained in paper 1 and the methodology developed and applied in paper 2, paper three investigates their applicability to district-heating consumption data.

Paper 4) **Evaluating Generalizability of Smart Meter Electricity Consumption Clusters**. This paper investigates the temporal stability of clustering solutions created using smart-meter electricity data. It develops a methodology for evaluating and quantifying cluster stability across time.

The aim of all four papers is to ensure consistency and a natural progression from initial identification of the current state of the art to application in Denmark and finally to propose improvements to the current literature. The contribution of each of the papers to the research objectives and the progression is outlined in Table 1.

Research Objectives	Theme	Paper 1	Paper 2	Paper 3	Paper 4
Q1	Consumption clustering	x	x	x	x
Q2	State of the art	x			
Q3, Q4	Clustering		x	x	x
Q3, Q5	Electricity Data		x		x
Q3, Q4, Q5	Method Development		x	x	x
Q4	District Heating Data			x	
Q5	Cluster Stability				x

Table 1 - Overall themes of the four papers and their relation to the research objectives.

1.5. Contributions

The four papers included in this dissertation contribute to the literature in the following five key respects:

Paper 1 conducts a comprehensive analysis of the concurrent state-of-the-art in smart-meter consumption clustering, conducted through a systematic review which identifies influential studies, methodologies, datasets and results. 2099 unique academic writings are included in the analysis, outlining the frequent application of common methods that are not sufficiently equipped to analyze data with temporal components.

Papers 2 and 3 identify the existence of temporal dependencies in smart-meter data. This is shown using autocorrelation and is identified for electricity and district-heating consumption data. The existence of temporal components in the smart-meter data is not surprising, as they represent repeated measurements from the same meter. Paper 1 shows that this has never been investigated in the literature on smart-meter clustering before, and hence has never before been exploited to improve the clustering. Papers 2 and 3 propose methods for remedying this gap.

Papers 2 and 3 further show how K-Means clustering described in 3.2 can be improved by careful preprocessing of the input data to enable the algorithm to handle temporal components in the data. This has been applied successfully to electricity and district-heating data. Data transformation was conducted using the autocorrelation features described in 3.5 and the wavelet features described in 3.6. The two transformations deliver radically different solutions. The wavelet clusters are comparable to the normalized clusters in creating identical clustering solutions and cluster compositions, but they compress data. Autocorrelation features are able to deliver finer-grained cluster solutions and significantly reduce data dimensionality.

Papers 2, 3 and 4 apply a novel adaptation of cross-validation for unsupervised learning. The adaptation uses the cluster validation indices (CVI) described in 3.3 as pseudo cross-validation to identify variability. This approach is an unsupervised equivalent to cross-validation for supervised learning. The implementation applied in papers 2, 3 and 4 reports the average, minimum, and maximum values for each CVI, analysis of the distribution can produce statistical confidence bands. The method is a framework for generalizing the cluster selection in K-Means.

Paper 4 develops a variance evaluation methodology to quantify the stability of the clusters over time. The method evaluates the ratio between the actual variances between two periods, and is described in 3.7. Through this measure it is possible to assess the stability of the clustering solution at different time periods. This methodology is not confined to application in smart-meter clustering but is general in its definition and applicability. Table 2 produces an overview of each paper's contribution to the literature.

Contribution	Paper 1	Paper 2	Paper 3	Paper 4
Systematic Review	X			
Pseudo-Cross Validation		X	X	X
Autocorrelation Features		X	X	X
Wavelet Features		X	X	
Varatio				X
Electricity Data Clustering		X		X
District Heating Data Clustering			X	

Table 2 – The four papers' contributions to the literature. Pseudo-Cross-Validation and Varatio are novel methods developed in the papers. The systematic review uncovers the state of the art in smart meter consumption clustering. Autocorrelation features and wavelet features represent novel application of methodology not used before in smart meter data clustering. The data analyzed introduce Danish smart meter data from electric and district heating utilities to the literature.

2. Smart Meters, Data and Software

This chapter briefly discusses smart meters as a product and concept for metering consumption, including the legislation on the rollout of smart meters. Data sets from SydEnergi (SE) and AffaldVarme Aarhus (AVA) analyzed in this thesis are described, along with description of the data cleaning and the software utilized for purposes of data processing and modelling.

2.1. Smart Meters and Smart Meter Data

Smart meters are digital replacements for the analog meters installed with end-users and are used to meter their energy consumption. Unlike smart metering, measuring consumption using analog meters requires twice-yearly manual readings by the utility company or the consumer. Smart meters are the digital counterparts of analog meters and are capable of reading and reporting consumption at very short time intervals, typically every fifteen to sixty minutes depending on energy type. The “smart” in smart meters is for the moment tied to the ability to read consumption at high frequencies and to transmit these readings automatically to the utility company, thus eliminating the need for manual readings.

Danish law state that by the end of 2020 all Danish electric utilities must ensure the installation of smart meters with all end-users [17]. The relevant act further lists the requirements for the meters’ ability to transmit the data to a third party and for the consumers to be able to access their consumption data. In addition, the recording frequency must be at least every fifteen minutes, and the data thus generated must be stored in a centralized data hub run by Energinet, the owner and operator of the Danish high-voltage transmission grid and gas net.

The European Commission has placed a high priority on promoting the installation of electricity smart meters across the European Union and projects the penetration of smart meters to reach 72% by 2020, with many member states exceeding 80% [18]. It is estimated that each meter delivers benefits worth of €309, distributed between production, distribution and consumption. In addition, the electricity smart-meter and smart-grid rollout can reduce emissions in the EU by between 3-9% and annual household consumption by similar amounts [19]. The European Commission expects that leveraging the potential of smart meters will help reduce energy waste and help the EU to reach its climate target.

Smart meters are anticipated to play an integral role in the reduction of energy waste in the European Union. Through vast amounts of consumption data, it is expected that consumption flexibility can be identified and that the data can help design new tariff structures to leverage this flexibility and thus reduce grid strain through peak shawing. The study of smart-meter data, especially for electricity, is a very active research field. Smart-meter data have been analyzed for clustering, resulting in disparate consumption profiles. Paper 1 provides an overview of smart-meter data analysis that has been conducted by both methodology and country.

This thesis is structured around the analysis of two distinct data sets supplied by AffaldVarme Aarhus, a municipality-owned district-heating producer and supplier, and by SydEnergi, an electricity utility company serving southern Denmark.

2.2. SydEnergi Electricity Smart Meter Data

The electricity utility company SydEnergi (SE) has provided a data set containing meter readings from more than 260,000 meters in its supply district from January to December 2011 inclusive. More than 220,000

meters record every fifteen minutes, while the remaining meters record hourly consumption. SydEnergi's distribution area covers southern Denmark, this being reflected in the data set, as all Danish consumer types are included, from individual street lights to households and large public and industrial consumers. In total there are eighteen types of households, not all of which are habitable year-round and 132 types of public and industrial buildings. Subsets of SydEnergi's data for the city of Esbjerg, focusing solely on households connected to district heating, are analyzed in papers 2 and 4. Each paper has different requirements regarding the data-cleaning process, of which it provides a detailed discussion. SydEnergi's smart-meter data are supplied at the atomic level and contains information about individual meters and locations. As these data are sensitive, the papers only show aggregated results in graphs and for individual anonymized meters. SydEnergi's smart-meter dataset includes per meter attributes, such as; postal code, road id, housing category etc. These attributes are utilized for the selection of specific housing types in papers 2 and 4 and are defined precisely therein. In Table 3 anonymized outputs from eight meters are shown to give a sense of the data.

	ELECTRICITY METER ID (ANONYMIZED)							
TIME STAMP	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
01-01-2011 00:00	0.31	0.33	1.17	0.24	0.11	1.14	3.25	10.1
01-01-2011 01:00	0.28	0.35	1.18	0.23	0.17	1.05	2.67	9.7
01-01-2011 02:00	0.34	0.36	1.34	0.25	0.16	1.06	0.85	8.1
01-01-2011 03:00	0.27	0.25	1.23	0.23	0.16	1.17	0.82	11.4
01-01-2011 04:00	0.24	0.36	1.22	0.26	0.15	1.15	0.81	11.7
01-01-2011 05:00	0.24	0.35	1.15	0.23	0.16	1.13	0.89	12.2
01-01-2011 06:00	0.27	0.24	1.16	0.24	0.21	1.22	2.66	13.4
01-01-2011 07:00	0.29	0.23	1.14	0.22	0.15	1.11	3.00	11.9
01-01-2011 08:00	0.84	0.25	1.13	0.23	0.22	1.13	2.76	15.7
01-01-2011 09:00	2.07	0.26	1.22	0.26	0.19	1.15	2.76	16.4

Table 3 – Anonymized output from eight electricity meters. Consumption is measured in kWh. The data has been anonymized such to ensure data privacy in compliance with GDPR

2.3. AffaldVarme Aarhus Heat Exchange Stations

The data provided by AffaldVarme Aarhus (AVA) contain smart meter readings from 53 heat-exchange stations (HX). Heat-exchange stations are the equivalents of electric transformation stations for district heating and link the transmission grid for water pressurized at 120^{o}C with the 80^{o}C distribution grid water. Heat-exchange stations supply an area with district heat, the smart-meter data readings being aggregated within the area serviced by the station.

The readings supplied by AVA contain hourly readings from January 2017, and the aggregated data cover thousands of individual household meters. The aggregation makes it impossible to identify individual households in the data, making the data non-sensitive in any respect. A subset containing data from January 2017 is analyzed in paper 3. Table 4 shows output of the AVA data with readings from the first seven hours of January 1st, 2017, for several heat-exchange stations.

	HEAT-EXCHANGE STATION ID							
TIME STAMP	**101**	**102**	**103**	**104**	**105**	**106**	**107**	**108**
01-01-2017 00:00	1.1	3.8	3.5	3.7	3.5	12.7	2.9	7.1
01-01-2017 01:00	1	4	3.4	3.8	3.6	12.5	3	7.3
01-01-2017 02:00	1	4	3.5	3.7	3.4	12.5	3	7.2
01-01-2017 03:00	1	3.8	3.6	3.8	3.4	12.5	3.1	7
01-01-2017 04:00	1.1	3.9	3.7	3.9	3.4	12.3	3.1	7.1
01-01-2017 05:00	1.1	3.9	3.5	3.8	3.3	12.4	2.9	7.4
01-01-2017 06:00	1	4.1	3.7	3.9	3.3	12.8	3	7.4

Table 4 - View of the district heating consumption data as supplied by AVA. The dataset includes the date in the format dd-mm-yyyy HH:MM. The Heat-Exchange Station ID is the unique identifier for each station. The value at each time stamp is MWh consumed energy for heating.

Prior to analysis, the data are cleaned to ensure analytical integrity. This process removes meters containing missing values and corrects outliers; the exact process is described in paper 3. Relevant for all types of meter data is the high dimensional nature of the data. Each time step represents one dimension, resulting in 96 dimensions per day with recordings every fifteen minutes.

2.4. Data Cleaning

The data supplied from AVA and SE was of almost pristine quality with very little data cleaning needed. This quality of the data I attribute to the fact that it is the exact same data used for billing and thus strict requirements are enforced to ensure correctness. At AVA and probably also SE they have workers looking at the data to ensure correctness. Despite this focus on ensuring data correctness, I did however; encounter some outlying data in the HX data from AVA. Paper 3 describes the process of correcting these. Also for SE data I did encounter few meter which exhibited flat consumption or other unexplainable traits. In all cases where data cleaning was needed, steps were taken to reduce correction bias due to the cleaning. For AVA data was it necessary to impute data as described in paper 3. In the subset selection of the SE data for paper 2 some meter were discarded because of undesirable traits relating to meters not working properly, details are described in the paper. The number of meter discarded for SE is described in paper 2 and represents a very small fraction of the entire dataset, and none were discarded for outlying values.

2.5. Software

This section briefly describes the software adopted for the modeling performed in the papers 2, 3 and 4.

The Python programming language version 3.6.x [36] was applied as base for the development of the different calculations needed to create the results for this thesis. The Python language was selected because of it ability for fast prototyping by alleviating the need for rigorous and stringent class and method declarations. The following Python packages were applied.

The data structuring and manipulation was primarily done using Pandas 0.20.1 [37]. For the numerical computations and matrix calculations, especially for calculating the cluster validation indices Numpy 1.12.1 [38], [39] was applied. For the Wavelet analysis PyWavelet 0.5.2 [40] was utilized, and the statsmodels 0.8.0 [41] package delivered the statistical confidence band for selection of significant coefficients.

SKlearn 0.18.1 [42] delivered the machine learning framework for doing K-Means clustering and various other models. Scipy 0.19.0 [43] was utilized as support library for Numpy when doing scientific

computations. Matplotlib 2.0.0 [44] delivered the framework for visualizing the results. Many packages applied are included in the python standard library.

3. Theoretical Background

This chapter will describe the theory adopted in all four papers. The section is structured such that methods that are prerequisites to other methods are introduced first. Section 3.1 starts by introducing and discussing statistical learning, which it links to the smart-meter data described in chapter 2. Then the K-Means clustering algorithm is introduced in section 3.2. The selection of the best solutions in K-Means is based on cluster validation indices (CVIs), which are described in section 3.3. A novel approach to cross-validation for unsupervised learning is presented in section 3.4. Autocorrelation features are introduced in section 3.5, with a brief subsequent discussion wavelet features in sections 3.6. In section 3.7 Varatio, a novel method for estimating stability of clusters by applying knowledge of variance structures is introduced. This chapter will mainly deal with statistical theory but concludes with a description of Okoli's method for systematic literature in section 3.8.

3.1. Statistical Learning: Clustering versus Classification

Identifying differences in consumption patterns by applying smart-meter data on electricity consumption is a delicate exercise. If the clusters are constructed exclusively from smart-meter consumption data without knowledge of the true underlying categories, unsupervised methods are required. Conversely, if any category information is available, mapping the smart-meter data to these categories requires supervised methods. This thesis distinguishes between classification and clustering as supervised versus unsupervised, the former including information about the actual categories of, for example, consumers, while the latter does not.

3.1.1. Supervised Classification

In Supervised classification**,** input data matrix $\underline{\underline{X}}$ is mapped onto a category vector $\vec{Y}$ using (non)-linear function(s). In this approach, regression analysis is one of the most widely applied methods. There also exists a well-developed framework for mapping and validation by means of residuals in order to evaluate the difference between observed and calculated categories.

In the case of supervised classification, knowledge is available about the actual underlying categories. In perfect classification this equates conceptually to:

$$Category = link\ function\ (data) \tag{1}$$

Usually there is a discrepancy between the observed and calculated categories, but it is still manageable to calculate the residuals between them. Much of the statistical framework addresses evaluation of the residuals obtained through regression. In mathematical form, supervised learning can be described as:

$$y_i = f(x_i) + \epsilon_i \tag{2}$$

where i denotes the i'th observation, and f can be any type of function, linear or non-linear, linking observations x_i in $\underline{\underline{X}}$ to the corresponding observation y_i in the response $\vec{Y}$. The term ϵ_i is the residual

accounting for the difference between observed and estimated. It is an unstructured error term where elements are identically and independently distributed, usually following the standard normal distribution.

3.1.2. Unsupervised Clustering

In Unsupervised clustering**,** input data matrix $\underline{\underline{X}}$ has no known corresponding mapping category vector $\vec{Y}$. Rather the data are split into k clusters, which by some similarity measure are evaluated as being more homogenous than the overall data $\underline{\underline{X}}$. The omission of categories for mapping exploited in supervised classification makes it difficult to directly validate the clustering produced in unsupervised clustering. This is attributable to the inability to calculate residuals as the discrepancy between the true categories and the estimated clusters. To circumvent the missing residuals, a large number of cluster validation indices (CVI) have been proposed and tested. The most prevalent CVI's in smart-meter data-clustering are identified in paper 1 and described in section 3.3. Prominent methods for unsupervised clustering include K-Means described in 3.2 and hierarchical clustering.

As unsupervised clustering does not include knowledge about the categories, we must identify some other measure capable of splitting the dataset into subsets, thereby creating more homogeneous clusters capable of identifying individual categories. Clustering is not easily expressed in one enclosed mathematical form, and several different measures have been developed. They employ different approaches to create splits in the data, often using a Euclidian distance measure. The distance measure chosen can heavily influence the resulting clusters [45].

It is impossible to evaluate residuals using unsupervised methods as there is no notion of difference between the categories and the clusters. Typically multiple clustering solutions are created, and the best is selected.

Needing measures to evaluate the clustering and to avoid the trivial case of assigning one cluster to each observation, thereby reducing the cluster variation to 0, a number of cluster validation indices are used. These indices evaluate the intra-cluster distance and relate it to the inter-cluster distance. Often the indices favor clustering solutions that minimize the intra-cluster distance while maximizing the inter-cluster distance. The concept is inspired by the residuals in supervised learning. This thesis has applied unsupervised learning to clustering smart-meter consumption data.

3.1.3. Temporal Components

Smart-meter data is recorded over time, potentially including a temporal component that can convey significant information about the consumption. The clustering techniques applied in smart-meter consumption clustering do not take this temporal component into account [35], consequently a very important feature of the data is not included in the clustering. The impact of a temporal component can be significant as shown in Figure 2, where the temporal component (right) enables the identification of structures not identifiable without (left).

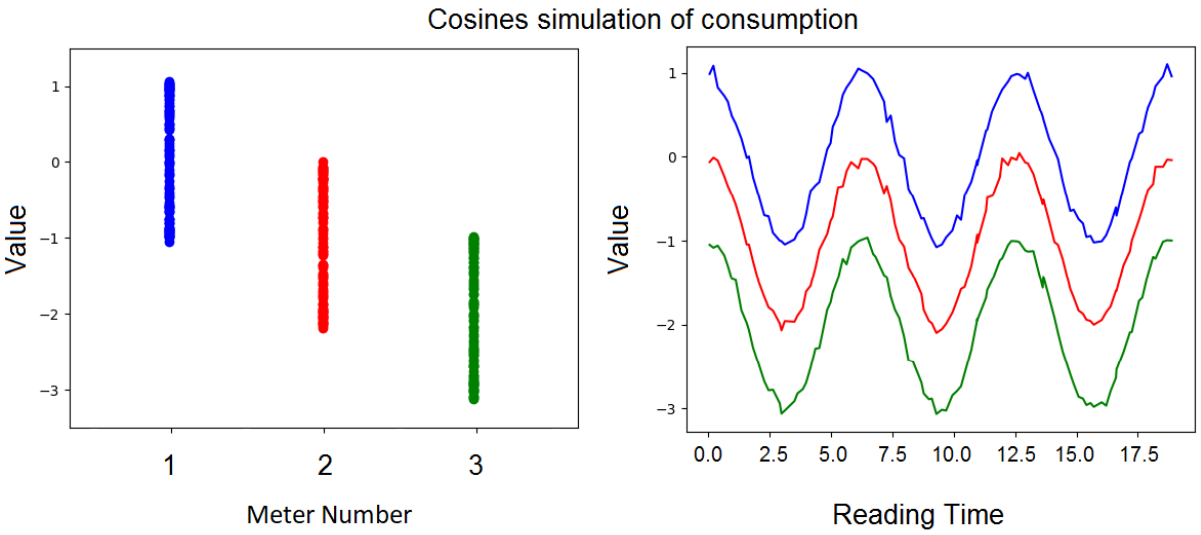


Figure 2 – Simulated temporal cosine components influence on data perception. Left: a scatter of points, collapsed to have no temporal component (Reading time). The three colors indicate three different clusters, but it is not possible to identify overlap. Right: the scatter has been expanded into its original temporal component. This graph is an adaptation from paper 2.

The cosine structure is clearly identified when the temporal structure in the data is analyzed revealing three different non-overlapping structures (right). It is also shown that none of the data overlap. Without this temporal information, it is not possible to determine whether the data overlap or are just very close in distance (left); the temporal component helps differentiate the data. The temporal component (right) is where the K-means and other unsupervised methods are hard-stressed, as they do not account for this when clustering.

Specifically, in the case of K-Means, the data will be evaluated at each time step independently of neighboring time steps, as shown in Figure 2 (left). Preprocessing the data before clustering can help K-Means to include the temporal structure, as discussed in papers 2 and 3.

3.2. K-Means Clustering Algorithm

The K-Means algorithm clusters data into homogeneous subsets, a method that is simple, efficient, robust and often able to successfully produce clustering results. Rather than breaking down when encountering unsupported data structures, it ignores intrinsic data structures it is unable to handle, such as autocorrelation, but creates clusters none the less. Its robustness can be interpreted as a “brute force” clustering approach as, regardless of data quality, it delivers a clustering solution in most cases. This robustness can also lead to unintended results if applied haphazardly. The robustness of K-Means should not be mistaken for stability of the solution.

The algorithm is described in Table 5 and consists of four steps: 1) initialization of the algorithm is done by randomly assigning k clusters to the data; 2) recursively each smart meter is assigned to the cluster closest “in distance”; 3) Each assignment updates the cluster means; and 4) steps 2 and 3 are repeated until there is no change in the assignment of clusters.

K-Means Clustering Algorithm [46]

1. **Randomly assign k = 1, 2, 3, …, K clusters, K is defined by the analyst.**

2. **For a given cluster assignment C, the total cluster variance**

$$\min_{C,\{m_k\}_1^K} \sum_{k=1}^{K} N_k \sum_{C(i)=k} \left||x_i - m_k|\right|^2 \quad (3)$$

   is minimized with respect to the means $\{\boldsymbol{m_1}, \dots, \boldsymbol{m_k}\}$ yielding the means of the currently assigned clusters. N is the number of observations, and $\boldsymbol{x_i}$ is the i'th observation vector.

3. **Given a current set of means $\{\boldsymbol{m_1}, \dots, \boldsymbol{m_k}\}$, (3) is minimized by assigning each observation to the closest (current) cluster mean.**

$$C(i) = \underset{1\le k\le K}{\operatorname{argmin}} \left||x_i - m_k|\right|^2 \quad (4)$$

4. **Steps 2 and 3 are iterated until the assignments do not change or the maximum number of iterations is reached. This algorithm can lead to suboptimal local solutions.**

Table 5 - K-Means algorithm.

K-Means is prone to delivering suboptimal solutions that can be unstable, as the method can get caught up in local optima due to the random initialization. Therefore, it is advisable to rerun the algorithm with different random initializations and subsequent selection of the preferred solution. The SKlearn package applied throughout the papers implements ten random initializations with subsequent selection of the best performance.

Apart from its random initialization, the K-Means algorithm is a deterministic algorithm whose objective is to minimize the distance from each observation to the cluster centroids. The centroids, which are average values of the members in the cluster, are updated each time a new member is added or removed. The constant updating of the centroids results in members leaving clusters and vice versa. This is continued until convergence is achieved, measured such that no member or centroid changes.

The K-Means algorithm evaluates each variable by itself, disregarding correlation information. For smart-meter data this equates to evaluating each metering time step independently of other time steps. In many settings this poses no problem for the clustering of the data, but for smart-meter data this has an effect. Smart-meter data as shown in 3.5 and papers 2 and 3 contain a time-dependent component shown by the existence of autocorrelation in the data. This component governs information about how previous consumption affects current consumption. As shown in Figure 3 K-Means evaluates each value on the x-axis (time step) independently, though the figure indicates periodicity. K-Means is unable to include this autocorrelation, and hence this information is not conveyed in the clustering solution. The inclusion of autocorrelation information could potentially decrease variability. Papers 1, 2 and 3 discuss the implications of excluding temporal information from the clustering.

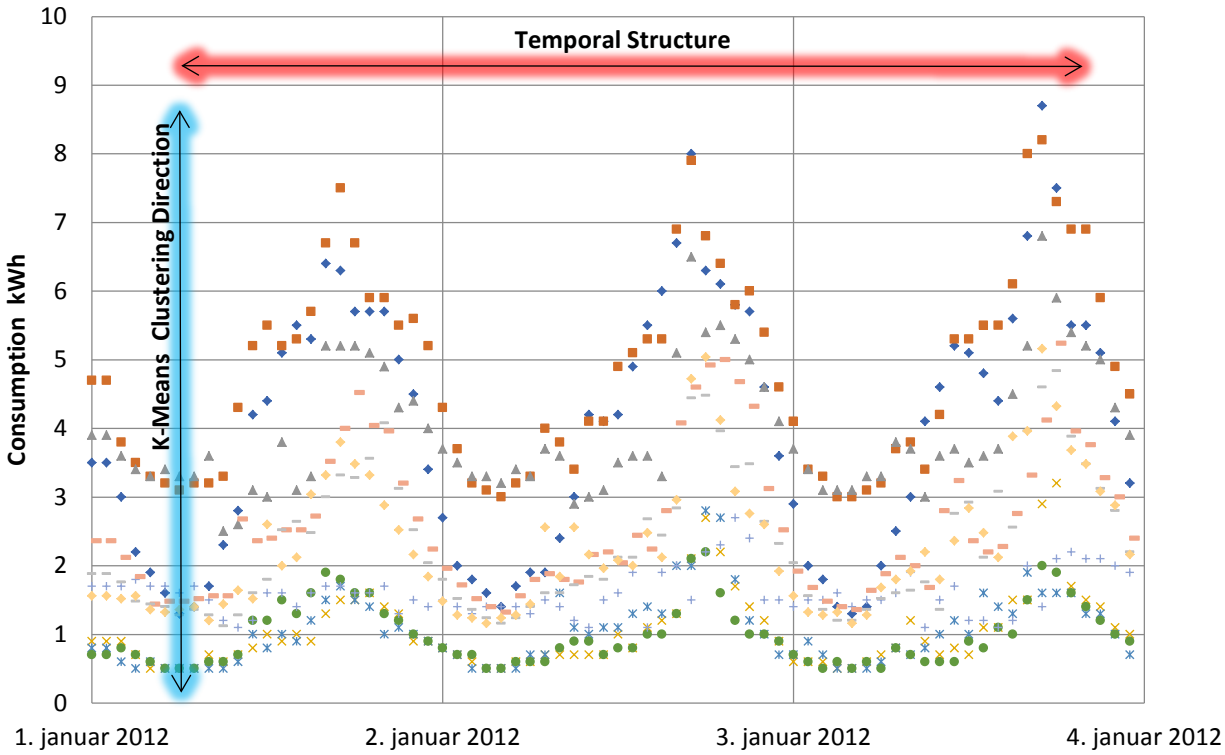


Figure 3 - Three day Scatter plot of 10 smart meters with hourly resolution. The blue direction arrow shows the direction of the K-Means computations when clustering; all observations in same hour are used for the clustering. Red direction arrow shows the temporal structure in the data, this structure is not included in the K-Means clustering.

As described in paper 1, K-Means is the most prevalent clustering algorithm in smart-meter consumption clustering, its simplicity and widespread availability makes it an obvious option for analysts. Paper 1 indicates that very few papers acknowledge the existence of autocorrelation in smart-meter data, and only one of the identified papers deploy time-series methods by applying Fourier transformation.

The simplicity of the K-Means algorithm makes it an excellent baseline for clustering. Papers 2 and 3 successfully investigate how careful preprocessing of the input data can enable K-Means to account for autocorrelation without changing the algorithm. It is possible to do this without introducing complexity in the clustering phase by transforming the input data such that the transformed data account for the dependencies. This enables K-Means to account for time dependencies indirectly, thereby including latent information and reducing the variance in the resulting clusters. The preprocessing of the data does not increase the computational cost, as the chosen transformations – autocorrelation feature and wavelet features – are calculated by applying efficient linear time algorithms [47].

The K-Means method is implemented in every major data science software package from proprietary to open source. Its simplicity makes it straightforward to implement, such that analysts can still deploy it even if it is missing from their preferred programming language.

The simplicity of the algorithm makes it easy to evaluate its computational cost. The evaluation of algorithms is done using O-notation, which evaluates upper bound computational cost by an order of magnitude [48]. The worst-case running time for the K-Means algorithm is $O(k^n)$ [49] for k clusters and n observations in the case of smart meters, n being number of meter readings and equating to dimensions in the dataset. The worst-case scenario is the maximum computational effort needed to cluster a given

dataset. The best possible running time for the K-Means algorithm is O($k^{\sqrt{n}}$) [49], a significant reduction of computational effort even for small datasets. In both upper and lower bound running times there is a significant speed gain to be harvested by reducing the number of observations per meter, e.g. the dimensions of the data. Some of the methods described in sections 3.5 and 3.6 have a significant impact on the running time. From papers 2 and 3 we have the following results from the K-Means clustering of different datasets. Table 6 shows the effect on electricity data and Table 7 on district-heating data. The two tables show that dimensionality reduction and data transformation by autocorrelation features and wavelets described in sections 3.5 and 3.6 significantly reduce the worst-case running time. It also has a positive effect on the best-case running time.

Processing (Electricity Data)	*Normalization*	*Autocorrelation Features*	*Wavelet Features*
Scaling / Transform	O(n)	O(n)	O(n)
Size of input data (n)	168 x 32k+	24 x 32k+	42 x 32k+
Best-case running time	$12^{\sqrt{168}}$	$12^{\sqrt{24}}$	$12^{\sqrt{42}}$
Worst-case running time	12^{168}	12^{24}	12^{42}

Table 6 - Runtime comparison table from paper 2. The Normalized and Wavelet methods were unable to provide meaningful clusters and are for comparison set to twelve clusters, and 25% compression for wavelets. The autocorrelation and Wavelet method reduce dataset size, with significant impact on the runtime. An adaptation from paper 2.

Processing (District Heating Data)	*Normalization*	*Autocorrelation Features*	*Wavelet Features*
Scaling / Transform	O(n)	O(n)	O(n)
Size of input data (n)	744 x 49	24 x 49	161 x 49
Best case running time	$4^{\sqrt{744}}$	$7^{\sqrt{24}}$	$4^{\sqrt{161}}$
Worst case running time	4^{744}	7^{24}	4^{161}

Table 7 - Runtime comparison table from paper 3. The different scaling and transformations identify different number of clusters in the data. In this case we can see that the worst-case running time for the autocorrelation feature clustering is better than the scaled or wavelet transformed data. An adaptation from paper 3.

The K-Means algorithm is sensitive regarding differences of scale between variables. Normalization of variables is often a requirement of meaningful clustering. Papers 2, 3 and 4 all employ some type of scaling or transformation. Paper 3 evaluates the four different scaling methods presented in Table 8 and their impact on the resulting clusters.

Scale	*Mathematical Description*	*Intuition*
Normalization	$\frac{x - x_{min}}{x_{max} - x_{min}}$	Normalization puts all observations on a 0-1 scale compared to the largest reading. Dimensionless.
Standardization	$\frac{x - x_{mean}}{\sigma}$	Standardization scales all observations compared to the standard deviation of the data. Dimensionless.
Mean-Center	$x - x_{mean}$	Mean-centering removes the mean from the meter reading. It is equal to shifting on the y-axis.
Mean-Divide	$\frac{x}{x_{mean}}$	Scales observations relative to the series mean. Does not constrain the y-axis to the interval [0, 1]. Dimensionless.

Table 8 - Scaling methods applied to K-Means input data. As presented in paper 3.

3.3. Cluster Validation Indices

Selecting the correct number of clusters for K-Means to calculate is no trivial task. K-Means is unable to assist in the selection of the number of clusters and only has the ability to calculate the clusters. To aid in the selection of clusters, several Cluster Validation Indices (CVI) have been developed. CVIs enable comparison and selection of the optimal number of clusters in the dataset. As there is no knowledge about the true underlying clusters, CVIs are unable to produce the exact number of clusters. However, by applying several CVIs in synergy, they can reveal information about the clusters produced by i.e. K-Means.

If the optimal number of clusters, denoted k*, is believed to lie interval K ranging from two and twenty, K-Means is applied to calculate cluster solutions for each number of potential clusters k in that range K. The CVIs are then also calculated for each number of clusters in K.

The CVIs for all clusters in K are plotted, and where there is a significant change in the development of a CVI, this indicates a potential k*. If more CVIs exhibit interesting structures around the same region, this emphasizes the region's importance. Usually the structure needed for a region to be interesting is identified by an "elbow break" structure, a sharp decline followed by a break and then by horizontal stabilization of the CVI values, looking much like an elbow, as illustrated in Figure 4.

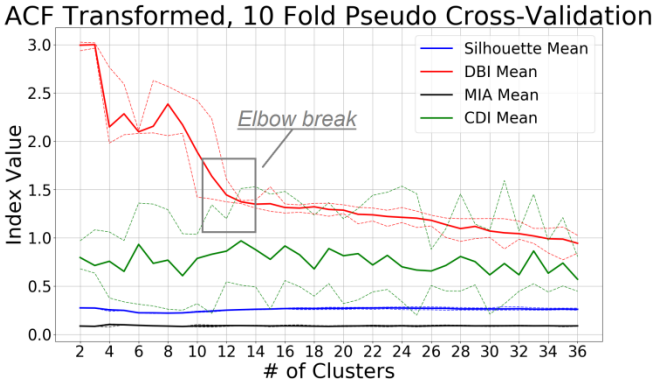


Figure 4 - Cluster validation indices developing as a function of clusters. The grey box marks a distinct elbow break indicating the optimum number of clusters. As presented in paper 2.

Paper 1 identified a large variety of different cluster validation indices (CVI) developed to evaluate the inter- and intra-cluster distances between the clusters. The most prevalent CVIs identified and presented in paper 1 are listed in Table 9. The table was also included in paper 1.

INDEX	MATHEMATICS	INTERPRETATION
DBI (DAVIES-BOULDIN INDICTOR)	$\frac{1}{K}\sum_{i=1}^{K}\max_{j\neq i}\frac{diam(C_i)+diam(C_j)}{d(C_i,C_j)}$	$diam(C_k)$ is the average diameter of a cluster, $d(C_i,C_j)$ is the distance between centroids, and K is the number of clusters. DBI relates the mean distance of each class to the distance to the closest class [50]. Smaller values of DBI imply that K-means clustering algorithms separate the dataset properly [51]
CDI (CLUSTER DISPERSION INDICATOR)	$\frac{1}{d(C)}\sqrt{\frac{1}{K}\sum_{k=1}^{K}d^2(C_k)}$	CDI prefers long inter-cluster distances and short intra-cluster distances [31]. Small values indicate good clustering. $d^2(C_k)$ is the squared average distance within cluster k (high), while $d(C)$ is max cluster distance in data.
DUNN	$\frac{\min d(C_i,C_j)}{\max diam(C_m)}$ where $d(C_i,C_j)=\min_{x\in C_i,y\in C_j}\lvert\lvert x-y\rvert\rvert$ and $diam(C_k)=\min_{x,y\in C_k}\lvert\lvert x-y\rvert\rvert$	The ratio between "minimum distance between clusters" and "maximum distance within clusters". When minimum dissimilarity between clusters increases in size and maximum cluster diameter becomes smaller, the Dunn value becomes large and indicates good separation. C_i is cluster i, d is distance, and m is total number of clusters.
SILHOUETTE	$\frac{c'(x)-c(x)}{\max\{c(x),c'(x)\}}$ $c'(x)=\min_{y\in C'}d(x,y)$	$c(x)$ is the average distance between vector x and all other vectors of the cluster c to which x belongs. $c'(x)$ is the minimum distance between vector x and all other vectors in cluster $\forall\ C'\neq C$ [51] SI is between [–1, 1]; higher is better. Negative is miss-clustering.
ENTROPY	$-\sum_{i=1}^{c}p(^i/_t)\cdot log_2 p(^i/_t)$	$p(^i/_t)$ denotes the proportion of correct classified vector i in cluster t. Entropy is a supervised index, as the true classes need to be known. Entropy is used as a measure of misclassification in each cluster. Entropy is small when the clustering result is similar to the expected result [31]. c is total clusters.
MIA (MEAN INDEX ADEQUACY)	$\sqrt{\frac{1}{K}\sum_{k=1}^{K}d^2(C_k)}$	Average distance within class to class centroid, summarized across all classes. k is the number of clusters; $d^2(C_k)$ is the squared average distance within cluster k. High MIA indicates large distances within the classes, e.g., large dispersion.

Table 9 - Cluster Validation Indices for finding optimum number of clusters in data, as presented in paper 1.

3.4. Pseudo Cross-validation

This section will discuss the pseudo cross-validation (PCV) methodology pioneered in this thesis for reducing bias when selecting number of clusters in K-Means and increase the generalizability of the solution. Pseudo cross-validation presents a general framework for evaluating generalizability when selecting optimal number of clusters for K-Means. It can be regarded as cross-validation for K-Means and is applicable for all types of data when using cluster validation indices. The method has successfully been applied in papers 2, 3, and 4 to help select the optimal number of clusters. Section 3.4.1 briefly describes cross-validation and motivates the need for this methodology. Section 3.4.2 describes the mechanics, while 3.4.3 outlines the algorithm and finally a discussion of the applicability and future improvements of the method in section 3.4.4.

3.4.1. Motivation

As discussed, in 3.2 the K-Means algorithm does not aid in selecting the optimal number of clusters, it simply generates the number of clusters k requested by the analyst. Cluster validation indices (CVI) have been developed and employed to help select the optimal number of clusters k*. The CVI's are calculated for all number of clusters k_j in the cluster interval K $= \{k_i, \dots, k_m\}$, for $i \leq j \leq m, \forall i, j, k \in Z_+$ before selecting the j giving the optimal clusters k*. The CVI's are point estimates calculated for each k_j and biased towards the data. The risk of overfitting K-Means by fitting the CVI's to the data is apparent and ultimately impairs the generalizability of the clustering.

In this thesis I introduce a novel method of reducing the bias of the CVI and improve the generalizability of the clustering solution selected. It does so by extending the concept of cross-validation outlined below, to K-Means clustering using any CVI metric. The method is not restricted to smart meter data and can be utilized whenever K-Means and CVI's are used for clustering.

When modelling a dataset, the model is intended to mimic the underlying data generating process, and not just the current realization. Increasingly complex models can improve the fit to the data, biasing the model towards the data, and impairing the generalizability of the model. It is a delicate process to avoid overfitting and optimally the model should be tested on a second dataset. Often this second dataset does not exist, so a true test dataset is unavailable.

Supervised learning, where a known response vector $\vec{Y}$ is modelled using a function f on the dataset matrix $\underline{\underline{X}}$, has introduced the concept of cross-validation (CV) to remedy the lack of a second dataset. It randomly divides the dataset into $Q = \{q_1, \dots, q_{max}\}$ roughly equal-sized mutually exclusive partitions and thereafter treating the partitions $Q \backslash q_i$ as a training set and the partition q_i as the test set, this is done for all $q_i \in Q$. It is assumed that the number of partitions Q is smaller or equal to the number of observations N, such that Q ≤ N. For each $Q \backslash q_i$ the response $\vec{Y}_{Q/q_i}$ is modelled, the model is then tested on the response $\vec{Y}_{q_i}$ from the remaining partition q_i. The process helps to quantify the model performance and to reduce the risk of overfitting. In turn all datasets are treated as test and training sets, resulting in a measure of variability within the dataset, the process is illustrated in Figure 5. The framework for cross-validation is not extended to unsupervised learning, but has to be developed case by case [52].

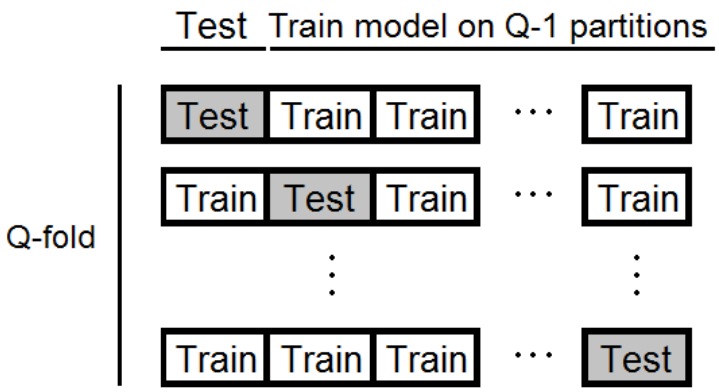


Figure 5 - Visual representation of Cross-validation. Q-fold shows the number of runs of modeling Q-1 subsets. The horizontal training boxes are applied for training the model while the gray "Test" box represents the test set which tests the model.

In this thesis K-Means has been selected as the clustering algorithm, due to its large prevalence in the literature. As it is impossible to identify the true underlying clusters in K-Means, the method is unsupervised and as such CV cannot be applied for model generalization.

3.4.2. Mechanics

In general, in supervised learning the prediction error estimate is defined as:

$$L(y, \hat{y}) = y - \hat{y} \tag{5}$$

Where $L(y, \hat{y})$ is the loss function, defined as the difference between the observed value y and the fitted value $\hat{y}$. By repeated modelling of partitions q of y and $\hat{y}$ through CV, a generalization of the loss is created. When the number of partitions Q is equal to number of observations N it is known as leave-one-out cross-validation in which the prediction error is an approximately unbiased estimate of the true prediction error [46]. The selection of Q is a trade-off between the computational effort needed and the bias-variance of the prediction error estimate. The cross-validation (CV) estimate of the prediction error is then:

$$CV = \frac{1}{|Q|} \sum_{q \in Q} L(y_q, f(X_{Q \setminus q})) \tag{6}$$

Where y_q are the observed values for partition q and $f(X_{Q \setminus q})$ are the estimates from partition Q\q. CV is the average prediction error across all partitions. In unsupervised learning we cannot define a loss function; pseudo cross-validation omits this by substituting the loss function $L(y, \hat{y})$ in (6) with any cluster validation index (CVI) and evaluating the development of the CVI. Using the CVI definition as loss function in (6), enables assessment of the variability of the CVI using the partition structure from cross-validation, thereby producing the $pseudo\_CV_j$ estimate for the j'th cluster k:

$$pseudo\_CV_j = \frac{1}{|Q|}\sum_{q\in Q} CVI(X_{Q\setminus q,}k_j) \quad (7)$$

Where $CVI(X_{Q\setminus q,}k_j)$ is the selected cluster validation index calculated for Q\q partition of the dataset X for k_j clusters. This function calculates the CVI for a specified number of clusters k_j in the interval K of interest, where $K = \{k_i, \dots, k_m\}$ for $i, m \in Z_+ \leq N$, and for all partitions of Q\q. As the CVI is calculated for all combinations of Q\q, its value will differ between Q\q for different q. The pseudo_CV is exploiting this fluctuation to reduce bias.

As the method does not validate the CVI itself but evaluates the variability of the CVI estimator it is not cross-validation per se, hence the estimator is called pseudo cross-validation. The methodology is similar to cross-validation, but the loss function is replaced with the CVI estimator. An inherent problem in unsupervised clustering is the lack of knowledge of the true categories we are trying to model; this means that it is not possible to select a prediction error estimator. Specifically, for K-Means the CVI's can be substitutes for the loss function $L(y, \hat{y})$, and thus produce an indirect evaluation of the difference between the modelled and the truth. The pseudo_CV reduces bias in the cluster selection process for K-Means clustering. This bias-variance trade-off increases overall generalizability of the solution. Pseudo_CV is applicable in cases where K-Means is selected for clustering and CVI's are used for selection of K.

3.4.3. Algorithm

The algorithm is similar in structure to cross-validation, substituting the standard loss function $L(y, \hat{y})$ with the cluster validation index $CVI(X_{Q\setminus q,}k_j)$. Below is a pseudo code representation of the algorithm for calculating pseudo_CV for K clusters.

Pseudo Cross-Validation Algorithm for K-Means

1) Randomly divide the dataset $\underline{\underline{X}}$ into q approximately equal-sized mutually exclusive partitions for a total of Q partitions.

2) For each number of clusters $k_j \; in \; K = \{k_i, \dots, k_m\}$, for $i \leq j \leq m$
3) For each partition q in Q do:

   a) Remove block q from the total dataset X
   b) Compute $CVI(\underline{\underline{X}}_{Q\setminus q,}k_j)$, from the remaining Q\q data partitions in $\underline{\underline{X}}$

4) The $pseudo\_CV$ is the average CVI across all blocks:

$$pseudo\_CV_j = \frac{1}{|Q|}\sum_{q\in Q} CVI(\underline{\underline{X}}_{Q\setminus q,}k_j) \quad (7)$$

5) Plot all $pseudo\_CV_j$ as function of j and select the optimal number of clusters k*.

Table 10 - Pseudo Cross-Validation Algorithm for K-Means

In papers 2, 3, and 4 the algorithm has been extended with minimum and maximum values for the each $pseudo\_CV_j$ visualizing the variability for each number of clusters k. Alternatively the distribution could be analyzed for distributional confidence bands.

3.4.4. Discussion and Applicability

The pseudo_CV reduces bias in the process of selecting the optimal number of clusters by evaluating variability in the cluster validation index (CVI). This variation in the CVI improves the generalizability of the selected solution. However, it does not improve the K-Means clustering solution itself. The method aids in the selection of the optimal number of clusters, not the specific cluster definitions created by K-Means. The pseudo_CV improves the generalizability and has been applied in papers 2, 3, and 4. Performance of CVI's has been investigated; this was done by evaluating the span of the CVI for each K. Some of the CVI's exhibited large variation such that it was inconclusive what number of clusters to select using specific CVI's, while others produced better estimates of the optimal.

The methodology of pseudo_CV is applicable to K-Means clustering whenever cluster validation indices are employed, and it is not tied specifically to smart meter data. Pseudo_CV presents a framework to alleviate bias in the cluster selection process for K-Means regardless of data, assuming CVI's are employed and the data is representative. In unsupervised learning where CV is not defined [52], pseudo_CV is an important contribution providing a framework for bias reduction when selecting the number of clusters for K-Means. It can be regarded as an unsupervised version of cross-validation for K-Means.

It is important to stress that the generalizability comes at a price, which is induced variance in the CVI estimate. For cross-validation it is well known that the process introduces large amounts of variance, this can make a solution look less favorable than is really the case. Given the similarity to CV there is no reason to assume pseudo_CV is immune to this inflated variance, but as the true values are unknown in unsupervised problems this property is difficult to investigate.

As development of and rigid investigation of pseudo_CV and its distributional properties was not a central aspect of this thesis, papers 2, 3, and 4 used the maximum and minimum values of the pseudo_CV. Further studies into the difference between the biased and pseudo_CV estimate of CVI is recommended, along with a more statistical rigid investigation of the distribution of CVI using pseudo_CV. This could potentially reveal interesting properties of how the biased CVI point estimate differs from the generalized CVI estimate. Furthermore, this could help evaluate if the CV property of approximately unbiased prediction error estimate for Q = N holds for pseudo_CV.

Pseudo_CV has not been tested for applicability to other clustering methodologies, conceptually if there is a metric used for evaluation, and then this metric can be evaluated using pseudo_CV. In an unsupervised setting indirect metrics is the closest to the truth we can come. Whereas supervised learning has many methods besides CV for estimating prediction error from dataset. To the best of my knowledge there is no similar method or alternative methods for reducing bias in the selection of optimal clusters k* for K-Means. For specific cases there might be custom methods which hold for that specific case. Such methods may well exist, but during the PhD-study I have not come across references to any such method.

3.5. Autocorrelation Features

Time Series analysis is the study of processes evolving over time. Smart-meter data as described in chapter 2 are time-series data. The smart-meters record consumption at equidistant time steps from the same household. The analysis of the smart-meter data conducted in this thesis does not completely embrace the rigor and structure of classical time-series analysis, in which the assumption is often that the time series is stationary.

A method from classical time-series analysis applied throughout papers 2, 3 and 4 is autocorrelation. Autocorrelation is a tool for identifying time dependency in data. It measures the time dependency between observations in the time-series data and identifies autocorrelation, if any, in the data. However, it does not remove autocorrelation, but rather quantifies it and determines whether it is statistically significant. In time-series analysis the autocorrelation is used to identify the specific underlying process describing a given time series. Here the autocorrelation is not used to identify a given process, as the interest is not in specific models, and therefore the rigorous assumption of stationarity of time series is not adhered to. The interest is in knowing which time lags τ are significant in describing the consumption profile. The sample autocorrelation function is defined as follows [53]:

$$ACF(\tau) = \frac{\sum_{t=1}^{n-\tau}(x_t - \bar{x})(x_{t-\tau} - \bar{x})}{\sum_{t=1}^{n}(x_t - \bar{x})^2} \quad (8)$$

where t and τ are integer time steps, $\bar{x}$ is the series mean. The structure reveals information concerning recurrence and periodicity in the smart-meter consumption data. A nice property of the autocorrelation is its invariance to scaling of data that is usually needed when applying K-Means clustering. Papers 2 and 3 identify autocorrelation in electricity and district-heating data respectively. Figure 6 shows the autocorrelation of a district heat-exchange station from the AVA data.

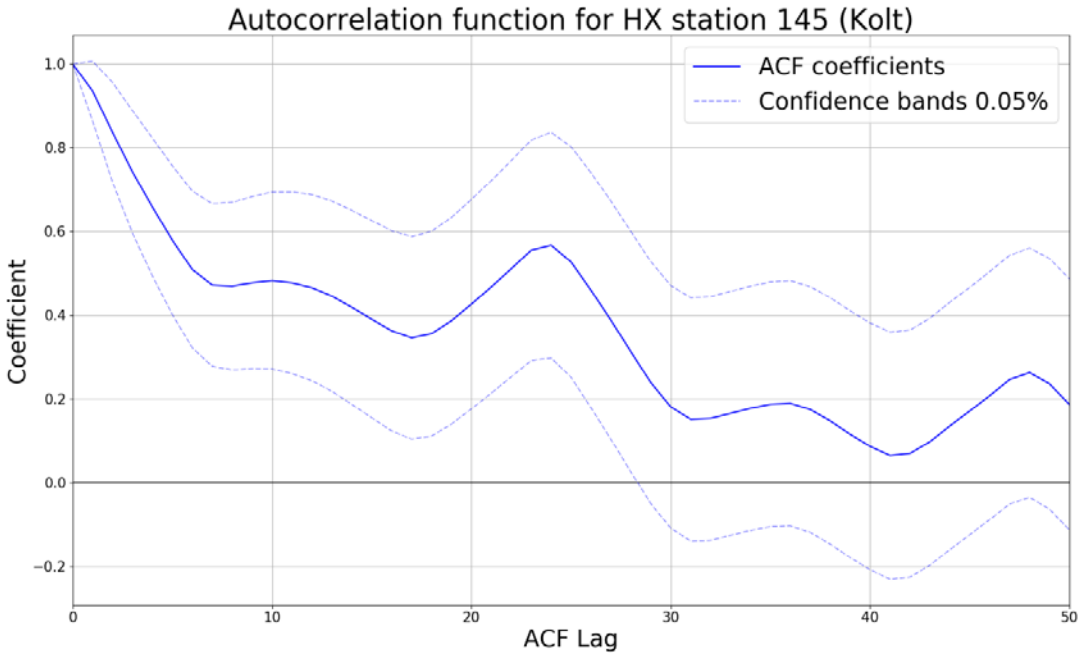


Figure 6 - Autocorrelation plot with 50 lags of heat exchange station 145 in the town of Kolt. The 0.05% confidence bands show significant lag coefficients until lag 28. A clear seasonality is also seen at lag 24, indicating a daily recurrent pattern. As presented in paper 3.

The main bulk of research on smart-meter clustering identified in paper 1 seldom acknowledges the temporal component. This thesis recognizes that smart-meter data is time-series data, and papers 2 and 3 demonstrate the potential existence of autocorrelation in smart-meter energy data. Figure 7 shows the significant autocorrelation coefficients for electricity clusters (left) and district heating clusters (right). Both smart-meter data types demonstrate autocorrelation, but the structure is radically different.

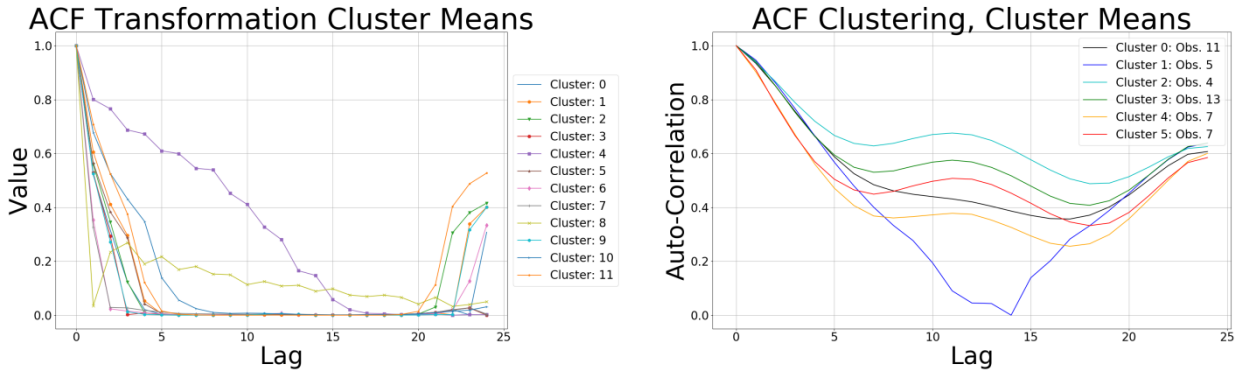


Figure 7 - Autocorrelation for electricity smart-meter clusters (left), autocorrelation for district-heating smart-meter clusters (right). Though both types of smart meter clusters exhibit significant autocorrelation, the structure is markedly different. Figures are from papers 2 and 3.

The literature review in paper 1 identifies only one paper applying time-series analysis techniques to the data [29] by applying Fourier transformation and subsequent clustering by the single largest frequency. Fourier transformation is a technique used in time-series analysis to transform the data from the time

domain to the frequency domain, but without keeping track of where specific frequencies are present in the former. Wavelet analysis, described in 3.6, can link time and frequency domain information.

3.6. Wavelet Features

Wavelets represent basis transformations through the application of wavelet basis functions. These functions are scaled and translated to fit the signal, that is, a smart-meter series. A notable property of the wavelet transformation is its ability to represent smooth and locally non-smooth functions through frequency and time localization, thus effectively linking time and frequency [46]. The ability to model local spikes and global smoothness makes wavelets appropriate for analyzing high-frequency data [54], enabling the wavelet to filter out high-frequency noise [47]. By filtering noise, wavelets efficiently compress the data, depending on the threshold selected by the analyst. In the case of the smart-meter data analyzed in papers 2 and 3, the compression factor was 5:1, but other studies, for example, in image analysis, have compression factors beyond 25:1 [55]. Figure 8 shows wavelet compression of smart-meter data as overlay to the original data.

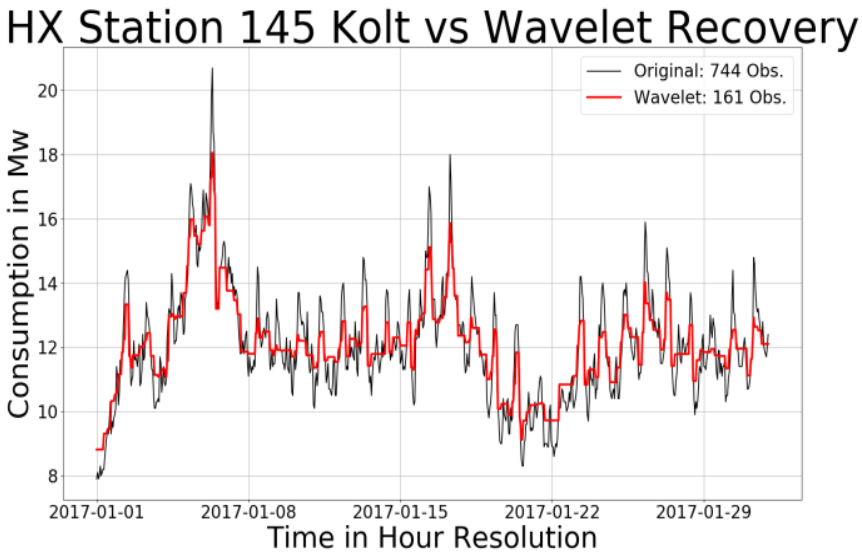


Figure 8 - Haar Wavelet approximation to the heat-exchange station Kolt; the compression factor is 5:1. The general structure of the consumption is captured by the wavelet, while the spikes are filtered out as noise. This selected wavelet has a large impact on the resulting fit, as does the threshold. As presented in paper 3.

The wavelet coefficients calculated through the wavelet transformation are not easily interpretable but are readily applicable as input to the K-Means. An important property of the resulting wavelet coefficients is they are uncorrelated [56]. The pyramid algorithm for computing the wavelet coefficient is very efficient and runs in $o(n)$ time [57], making it computationally feasible as a transformation for correlated smart-meter data. Further discussion of wavelets can be found in paper 3.

3.7. Varatio

This section will describe the Varatio methodology developed and successfully applied in paper 4 it enables easy comparison of clusters, of the same meters but at different periods in time, to evaluate the cluster stability across time. This chapter is split into three sections, in which section 3.7.1 describes the motivation for the methodology; section 3.7.2 describes the mechanics of the method, while section 3.7.3 discusses the applicability and prospects of the Varatio methodology.

3.7.1. Motivation

Varatio is developed as a tool for comparing stability of cluster solutions. In smart meter consumption clustering, this relates to identifying if clusters are identical across time i.e. weeks. Identical in this context means all members in one cluster in week 1, are consistently grouped together into the same cluster at later weeks and without inclusion of new cluster members. If clusters become less identical over time, it means the original members of the cluster are transitioning to other clusters and/or vice versa. Depending on the length of time, and the nature of the data, this may or may not be expected e.g. for the Danish smart meter data, I would expect the clusters to be almost identical when evaluating clusters of weeks back to back, but that clusters are less identical when comparing weeks very far apart. It is important to note that the cluster profile itself is dependent on the profiles of the members, meaning that if the members change profile then so does the cluster. E.g. for the district heating consumption, it is expected that the profiles of all members of a cluster change over the course of the year as the seasons change, but the clusters would encapsulate the same members.

Unless clusters are stable across time, we cannot make statements about the members within. If cluster definitions are unstable then the clustering must be redone every time a utility wants to select certain profiles. If a clusters composition changes only slightly over time, analyzing which members remain, and which leave the cluster can reveal valuable insights into how consumption patterns change over time.

Studying cluster stability requires consumption readings from various recording periods from the same meter. Many studies identified in paper 1 did not have access to datasets with multiple periods. The AVA and SE dataset analyzed in this thesis include recordings from an entire year, making it possible to analyze if meter profiles change throughout the year, and if transitions between clusters occur. In paper 2, the SE electricity consumption data is used to cluster profiles using data from the second week of January. This week was randomly selected, leading naturally to the question: would the cluster solution be identical if a different week was selected? This is what Varatio is developed to analyze and quantify.

Comparison of two clustering solutions can be written in tabulated form as seen in Table 11, with the week 1 cluster solution as rows, and the week 2 cluster solution as columns. If, in this notation, the solutions forms a diagonal matrix, then the clusters are identical (left), otherwise the clusters are not identical. As there is no apparent data structure for comparing several of these solution comparison matrices, the manual task of comparing more than two clustering solutions can be cumbersome. Varatio, developed in paper 4, measures the degree to which the comparison matrix is a diagonal matrix. It exploits general properties of the definition of sample variance to compare clusters from different clustering solutions, e.g. solutions from different weeks.

Best case mapping 1:1

Week 1 \ Week 2 Cluster	1	2	3	4	5
1	50	0	0	0	0
2	0	50	0	0	0
3	0	0	50	0	0
4	0	0	0	50	0
5	0	0	0	0	50

Worst case mapping 1:5 (1:k)

Week 1 \ Week 2 Cluster	1	2	3	4	5
1	10	10	10	10	10
2	10	10	10	10	10
3	10	10	10	10	10
4	10	10	10	10	10
5	10	10	10	10	10

Table 11 - Two extreme cases of weekly cluster mapping. Left shows the 1:1 mapping between two clustering solutions. None of the clusters from week 1 have members in multiple clusters of week 2. Right shows the worst case mapping 1:k, k=5; the members of all the clusters in week 1 are mapped uniformly to all clusters in week 2

Comparison of weekly cluster solutions for an entire year produces large amounts of such matrices. Paper 4 divides one year into quarters of approximately 12 weeks, each quarter is evaluated separately. This results in comparison between all 12 weeks of the quarter, giving 12*12 pairs of 2-dimensional comparison matrices. 12 occasions of the matrices are weeks compared to themselves and are subtracted for a total of 12*12-12 = 132 2-dimensional comparison matrices. Varatio can summarize the information of all 132 comparison matrices into 1 matrix.

3.7.2. Mechanics

When comparing clusters there are two possible types of outcomes illustrated in Table 11, A) the clusters are identical, having a 1:1 mapping or B) members of the clusters in one solution are scattered across clusters in the other solution, for a 1:k mapping, where k is the number of clusters in the second solution. For stability in a cluster solution, a 1:1 mapping is desirable and the Varatio metric can evaluate how close to such a 1:1 mapping, the clustering solution is.

Example 1: Consider two weeks of recording of hourly consumption data, collected from a group of smart meters. A K-Means clustering solution is produced separately, for each week, and both produce k = 5 clusters, week 1: $w1 = \{c_1, c_2, c_3, c_4, c_5\}$ and week 2: $w2 = \{c'_1, c'_2, c'_3, c'_4, c'_5\}$. The solutions w1 and w2 are compared to identify if smart meters remain in identical clusters across weeks. The cluster solutions are either A) identical, and for simplicity the labels too are identical, $\{c_1 = c'_1,\ c_2 = c'_2,\ c_3 = c'_3,\ c_4 = c'_4,\ c_5 = c'_5\}$, or B) the solutions are dissimilar, such that the cluster members in w1 are scattered across the clusters in w2, e.g. 25 percent of the members of c_1 are mapped to c'_1, 10 percent to c'_2 and the rest to c'_4 and so on. ■

Figure 9 illustrates the generalized case of example 1 and shows the difference between a 1:1 cluster mapping (left) where the members of a cluster in week 1 will be grouped together in week 2. This mapping presents stable clusters between the periods. Conversely, 1:k cluster mapping (right) illustrates the case in which the clusters in week 1 are scattered across the k clusters of the solution for week 2. It is important to notice that clusters can be identical in size between the two weeks, but their composition may be radically different, indicating unstable clusters. This means that the size of clusters between clustering periods is not a stable measure of cluster stability.

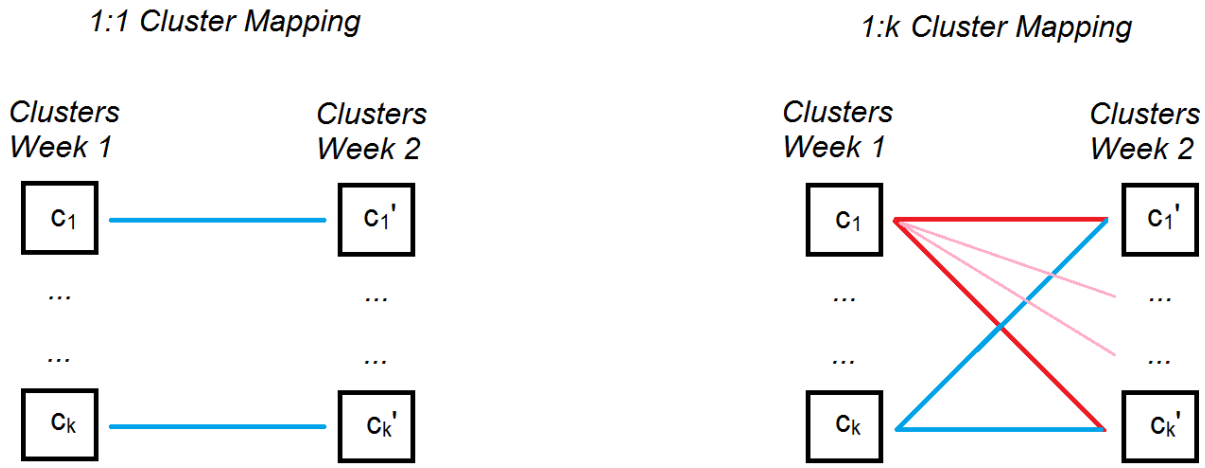


Figure 9 - Illustration of 1:1 cluster mapping (left) where clusters are stable over time. 1:k cluster mapping (right) indicates non-stable clusters.

The labels in K-Means cluster solutions are not persistent, meaning, rerunning the clustering can append new labels to the clusters even when the members are identical, as illustrated in Table 12. This property prevents cluster labels from being used for comparison of the clusters. Neither cluster size nor labels are an applicable metric for comparison of clusters, and instead each member in a cluster solution must be tracked to analyze stability of and transitioning of meters between clusters.

1:1 Mapping Non-Persistent Labels

Week 1 \ Week 2 Cluster	1	2	3	4	5
1	0	0	50	0	0
2	0	50	0	0	0
3	0	0	0	0	50
4	50	0	0	0	0
5	0	0	0	50	0

1:1 Mapping Rearranged Labels

Week 1 \ Week 2 Cluster	3	2	5	1	4
1	50	0	0	0	0
2	0	50	0	0	0
3	0	0	50	0	0
4	0	0	0	50	0
5	0	0	0	0	50

Table 12 - Clustering solution of two weeks with 1:1 mapping with non-persistent labels (Left).Reaaranging the labels producees a diagonal matrix revealing 1:1 mapping (Right).

The Varatio exploits the properties of the sample variance of random variables, to produce a metric which can quantify the difference between two clustering solutions into a vector or single scalar. That is, Varatio can reduce each of the 2-dimensional cluster comparison matrices into a vector which can be compressed into a scalar. The scalar is simply the average value of the Varatio vector. The Varatio vector shows how close to a 1:1 mapping each cluster in a solution is, while the scalar produces an average estimate of how close to 1:1 mapping the entire cluster solution is. As vector or scalar Varatio still conveys the information of similarity of clusters or solutions, but it is not possible to reproduce the 2-dimensional comparison matrix from the Varatio metric.

Varatio repeatedly applies the definition of sample variance:

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 \tag{9}$$

to the data. Where x_i denotes the i'th observation and $\bar{x}$ is the sample average. Since Varatio is developed for evaluating the stability of the clustering solution, there can never be a negative number of members in a cluster. During comparison as in Table 11 and Table 12, there can however be empty clusters due to the mapping. The sample variance for vectors of non-negative elements (size of clusters) is subject to two extreme cases: A) the minimum variance $s^2_{min} = 0$ is achieved when $x_i = \bar{x},\ \forall i$, and B) the maximum variance s^2_{max} occurs when all but one $x_i = 0$ in (9). It is important to note that the assumptions Varatio makes on the sample variance only hold for non-negative values of the cluster sizes. That is the s^2_{min} represents the 1:k uniform distribution of members among clusters, and s^2_{max} is when all members of a cluster are mapped 1:1 to another cluster.

Example 2: Consider 50 households grouped together in five groups. If they are distributed uniformly this can be illustrated as a vector = (10, 10, 10, 10, 10). The mean and sample variance of this vector are:

$$\bar{x} = 10,\ s^2 = \frac{1}{4}((10-10)^2 + (10-10)^2 + (10-10)^2 + (10-10)^2 + (10-10)^2) = 0.$$

If all the households from the 3rd group are then moved to the 5th group instead, producing the vector = (10, 10, 0, 10, 20), the mean $\bar{x} = 10$ remains the same, but the variance changes:

$$s^2 = \frac{1}{4}((10-10)^2 + (10-10)^2 + (0-10)^2 + (10-10)^2 + (20-10)^2) = 50.$$

Rearranging the meters in one cluster creates a vector = (50, 0, 0, 0, 0) the mean again remains $\bar{x} = 10$, but the variance now reaches the largest possible value for this constellation:

$$s^2 = \frac{1}{4}((50-10)^2 + (0-10)^2 + (0-10)^2 + (0-10)^2 + (0-10)^2) = 500.$$

This knowledge can be applied to evaluate the mapping between clusters. ■

Let us assume these 50 households, in example 2, where originally chosen because they had the same electricity consumption patter in the first week, w1 $= (c_1, c_2, c_3, c_4, c_4)$, $c_1 = 50$. Now assume each of the vectors: (10, 10, 10, 10, 10), (10, 10, 0, 10, 20), and (50, 0, 0, 0, 0) from example 2 are possible mappings of c_1 to week 2, w2 $= (c_1', c_2', c_3', c_4', c_5')$. The last vector (50, 0, 0, 0, 0), is a 1:1 mapping where all 50 household are mapped into a single identical cluster out of the five, c_1': $mapping_{1:1} = ({c_1}' = 50,\ {c_2}' = 0,\ {c_3}' = 0,\ {c_4}' = 0,\ {c_5}' = 0)$, representing a 1:1 mapping of c_1 into w2. This constellation produces the largest possible sample variance $s^2_{max} = 500$. If conversely the mapping of c_1 is 1:5 and uniformly distributed across the five clusters of week 2, just as the first vector in example 2, then the $mapping_{1:5} = ({c_1}' = 10,\ {c_2}' = 10,\ {c_3}' = 10,\ {c_4}' = 10,\ {c_5}' = 10)$ has zero variance and thereby the smallest possible variance for the mapping, indicating that c_1 is an unstable cluster. Table 13 shows the mean and variance for a 1:1 and a uniform 1:k, k=5 mapping for cluster 1 of week 1. The Uniform 1:k mapping results in zero variance, the smallest amount of variance obtainable, while 1:1 generates the largest possible variance with the data.

MAPPING OF CLUSTER 1 IN WEEK 1 TO CLUSTERS IN WEEK 2.		
CLUSTER NUMBER IN WEEK TWO	1:1 mapping s^2_{max}	Uniform 1: k mapping s^2_{min} (k=5)
1'	50	10
2'	0	10
3'	0	10
4'	0	10
5'	0	10
MEAN	10	10
VARIANCE	500	0

Table 13 - Two extreme cluster composition examples, uniform mapping and 1:1 mapping. Exemplified with 50 smart-meters from cluster 1 in week 1, distributed across clusters in week 2. There is a difference in the variance of the clustering; this variance difference is exploited in the Varatio measure.

The sample variances $s^2_{min}$and s^2_{max} constitute the two extreme boundaries of the variance for cluster mappings, with s^2_{min} representing the most scattering mapping 1:k and s^2_{max} the perfect 1:1 mapping. All other constellations of the variance $s^2_{observed\ mapping}$ are within these bound. Being able to calculate the upper bound of the variance makes it possible to evaluate how close to a 1:1 mapping, two clustering solutions are. If the observed variance divided by the maximum possible variance equals one (=100%), the clusters are stable. The larger the ratio, the closer the mapping is to being 1:1. The lowest possible ratio is 0, in which case $s^2_{observed\ mapping}$ = s^2_{min} and the mapping is uniform. The Varatio for evaluation of cluster stability is defined as:

$$Varatio = \frac{s^2_{observed\ mapping}}{s^2_{max,1:1\ mapping}} \tag{10}$$

With s^2 as defined in (9) and $x_i \geq 0, \forall i$. Varatio is defined as the ratio of the observed cluster variance and maximum obtainable cluster variance of the mapping. Since the number of cluster members is always known, when evaluating cluster stability, we can always calculate the maximum variance for each cluster mapping, by setting all but one cluster to size zero, as illustrated in Table 13. Varatio must be calculated for all combinations of clusters between the periods evaluated, creating the Varatio vector. The resulting vector must be evaluated by the analyst, as Varatio makes no assumption of what its vector or scalar should be and therefore must be evaluated on a case by case basis. A Varatio of 20% can desirable in some cases while in other cases the goal is closer to 100%. Paper 4 implements Varatio and applies it to the SE data for several different time periods. Figure 10 shows three possible calculations of Varatio for two weeks, each with five clusters containing fifty members.

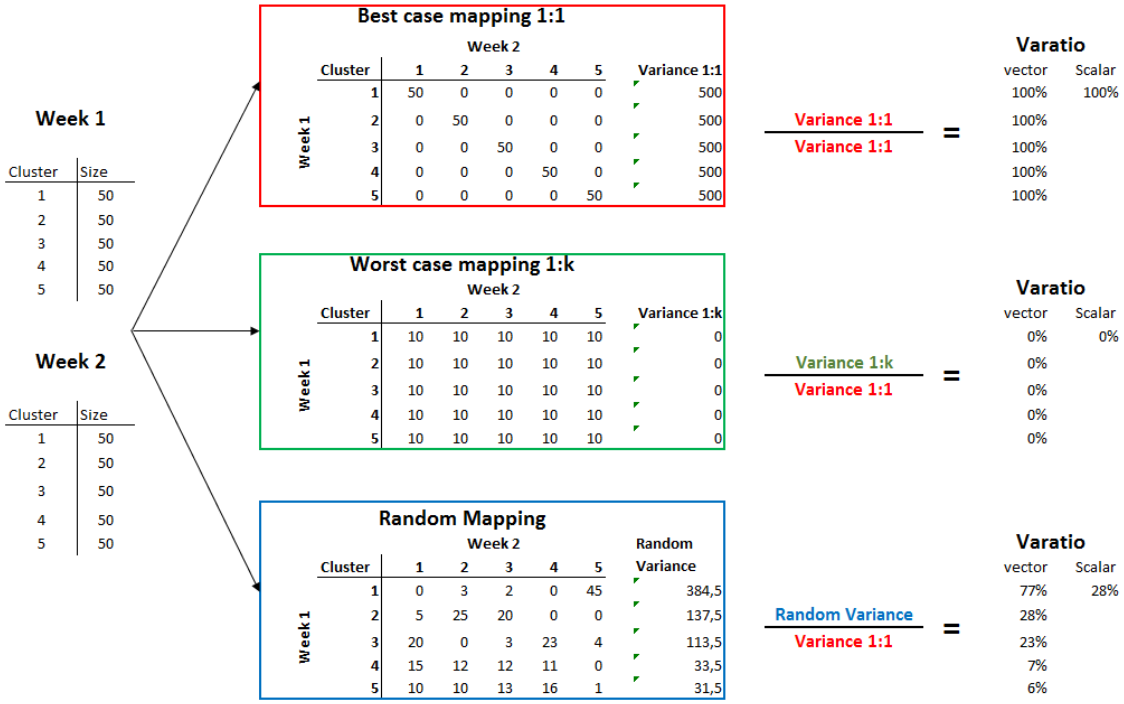


Figure 10 - Two weeks each with 5 clusters and 50 members in each cluster are subjected to the Varatio calculations. Best-case mapping appears in the red box, showing that clusters in week 1 are mapped 1:1 into clusters in week 2, resulting in Varatio of 100%. Worst-case mapping is uniform 1:k, shown in the green box – this is mapping the clusters from week 1 uniformly across clusters in week 2, resulting in Varatio of 0%. Finally the blue box random mapping shows how Varatio develops for different mappings from 1:1 to 1:k approximately. As presented in Paper 4.

3.7.3. Discussion

Using statistical variance of cluster changes, the Varatio metric can calculate how stable clusters are across time. It does so by compressing each cluster solution to a single scalar, which is a ratio of how close the solution is to stability. In paper 4, twelve weeks were compared at a time. Varatio reduced each of the twelve week cluster solutions into a vector and thereby reducing the entire problem from 132 2-dimensional matrices of clusters to a single 2-dimensional mapping matrix.

Varatio does not present new information when evaluating solution stability. It offers a new perspective on comparison by extracting the essential, namely how many members of a cluster continue to be clustered together across time. The information is available by examination of the comparison matrices. But like methods such as principal component analysis which through eigenvalue analysis highlights structures, Varatio uses variance and presents a more convenient overview for comparing matrices.

The method does not aid in the selection of clusters but can aid in the evaluation of the cluster solutions. As such it is a tool for evaluating if clusters created are a snap-shot of that specific period investigated or if the solution is stable, i.e. are identical clusters encountered at another time period.

Varatio does not interfere with any part of the clustering such as the selection of clusters, or the identification of clusters. It is a method which is applied after the clustering solution has been created and presents an overview of comparison of multiple clusters at multiple time periods. This also implies that regardless of change in the cluster profiles, the Varatio metric is unaffected as it evaluates the transition of every member of every cluster and is independent of the profiles.

3.8. Okoli's Systematic Review

Paper 1 conducts a systematic literature review of the current state of the art in consumption clustering using smart-meter data. Okoli [58] argues for a systematic process in literature reviews to produce a structured identification of the relevant literature, thereby enabling reproducibility of the study and the results. He argues that, even though a literature review is a pillar of every research endeavor, there is a lack of structure in how literature reviews are conducted except in dedicated review papers. The literature review conducted in paper 1 applies a modified version of Okoli's [58] process for systematic literature review, accommodating seven steps rather than the original eight. The methodology was originally developed for information science, where qualitative data extraction and subsequent quality appraisal are needed; this is not relevant for the quantitative sciences. Hence the steps *Extract Data* and *Appraise Quality* are merged into one step in paper 1. The seven steps of the modified method are as follows (and as stated in paper 1):

Modified Okoli Process for Systematic Literature Review
1. **Purpose of the literature review. Clearly state the purpose of the review. What is the scope and contribution of the work presented?**
2. **Protocol and training. Ensure consistency, alignment and reproducibility by formally defining rules and evaluation criteria.**
3. **Searching for literature. Explicitly describe the literature search, the "what and where."**
4. **Practical screen. Crude inclusion and exclusion of articles not based on quality appraisal but on "applicability to the research question." The reviewer normally only reads the title and abstract at this stage. "The practical screen is to screen articles for inclusion. If the reviewer thinks that an article matches the superficial qualities of the practical screen it should be included"** [58]**. If in doubt also, the article should be included.**
5. **Quality appraisal. Screen for exclusion, and explicitly define the criteria for judging articles. All articles need to be read and scored for their quality, depending on the research methodologies employed by the articles** [58]**.**
6. **Data extraction and synthesis of studies. Systematically extract the applicable information from the identified articles and combine the facts.**
7. **Writing the review.**

Table 14 - Modified Okoli method for systematic literature review. The method is presented in paper 1.

4. Paper Presentation and Results

This section presents the four research papers produced during the Ph.D. project, and summarizes the results and most important findings of each paper. The papers are presented individually so that their individual concepts and contributions may be clearly outlined. They are also organized in their intended progression and are each described in terms of: 1) scientific outline of the paper; 2) methodology used; 3) results; and 4) conclusions. The contributions of the papers are summarized in section 1.5.

4.1. Paper 1 - Structured Literature Review of Electricity Consumption Classification Using Smart-Meter Data

This paper has been published in the MDPI journal *Energies*, an open-access journal.

4.1.1. Scientific Outline

Paper 1 examines the current trends and research focus in the field of energy consumption clustering using smart-meter data. It applies a modified version of Okoli's [58] method to generate a systematic literature review ensuring the structure and reproducibility of the entire process. The data were collected from Thomson-Reuters Web of Science academic search engine, which indexes academic literature from books, conferences, symposiums and journal papers. It has indexed more than a billion academic texts, and searches across more than 12.000 journals [59].

4.1.2. Methodology

The paper conducts a systematic literature review to establish the state of the art in smart-meter consumption clustering using smart-meter data. It applies a modified version of Okoli's process for systematic literature review described in section 3.8. For feasibility the study only evaluates peer-reviewed journal papers indexed by the Thomson-Reuters Web of Science search engine [59].

The search engine enables keyword or search-phrase searches across multiple attributes. This paper utilizes this feature by searching for search phrases in the title and content of the literature. The attributes are applied in an "or" clause, enabling the search phrases to refer to either title or content. The paper used thirty search phrases relating to smart-meter data analytics, the search phrases being listed in appendix A of paper 1. Data collection was undertaken from July 5th to 12th 2016 and resulted in 3922 pieces of academic writing, of which 2099 were unique academic texts.

The systematic process commences by screening titles for inclusion, including all papers of potential relevance by adopting a "when in doubt include" criterion. This part includes papers which are later excluded due to their insufficient relevance. This process prunes the number of relevant papers to 552. Following the screening of titles, abstracts are evaluated, and non-peer-reviewed papers are removed. The 552 papers are subdivided into themes identified from the abstract. This results in ten distinct categories shown in Figure 11, and enabling an overview of research topics that use smart-meter data.

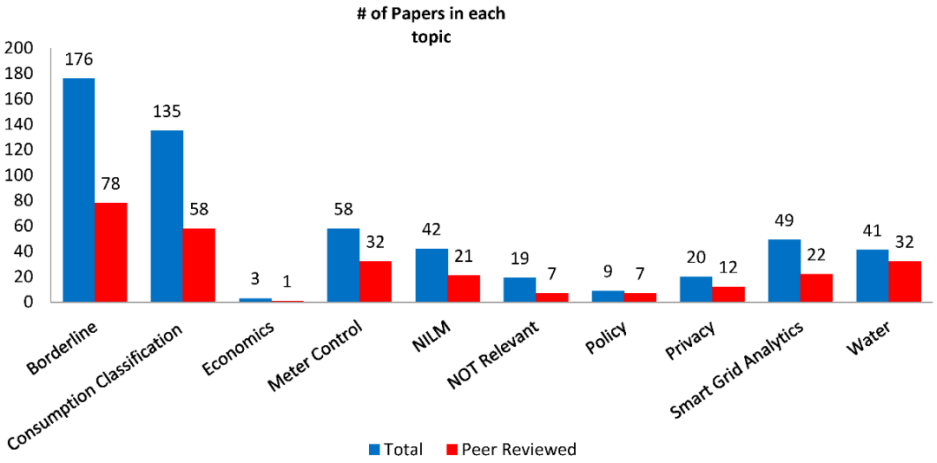


Figure 11 - Category distribution after abstract screening. The total includes all types of written material from papers, conferences, workshops etc. Peer-reviewed only covers journal papers. As presented in paper 1.

In the figure two categories are of interest regarding consumption clustering; Consumption Classification and Borderline. Consumption Classification encompasses all papers with abstracts describing the analysis of smart-meter data for purposes of classifying consumption. Borderline consists of papers where the abstract does not reveal whether smart-meter data were used for consumption classification and where no obvious fit to any of the nine remaining categories could be established. Papers in Borderline undergo extensive screening, addressing the data description and methodology used in order to evaluate their relevance to the classification of consumption. Only thirteen papers in Borderline are identified as relevant, resulting in a total of 71 peer-reviewed papers split between 58 in Consumption Classification and 13 in Borderline. At the time of the review clustering of smart-meter consumption had only been performed on electricity data. After its publication few papers clustering district heating consumption have been published [60], [61].

The final screening requires the 71 papers to be read to establish their relevance to consumption clustering. The process identifies 34 papers from which to extract and quantify relevant information. The effects on the paper bulk exerted by the individual screening processes are summarized in the waterfall statistic shown in Table 15. The column *Screening Type* describes the screening applied, *Bulk* indicates the remaining papers after application of the screening and *Reduced* quantifies the effect of each screening type.

Screening Type	Bulk	Reduced
Initial	3922	-
Unique	2099	1823
Screening I: Title	552	1271
Screening II: Abstract	311	241
Removal of non-peer-reviewed papers	136	175
Screening III: Borderline revisited	71	65
Screening IV: Reading of articles	34	37
Final number of papers synthesized	34	-

Table 15 - Waterfall statistics showing how many articles were excluded in each step of the screening process. As presented in paper 1.

4.1.3. Results

The paper identifies more than ten different methods for consumption clustering using smart-meter data. The most prevalent methods are K-Means, used in 65% of all papers assessed, and Hierarchical Clustering,

used in 29%. Only one paper uses time-series methods by employing Fourier transformation. The review also identifies eighteen different cluster-validation indices (CVI) for evaluating the clustering results thus obtained. The most prevalent CVIs are the Davies-Bouldin Index (DBI), the Cluster Dispersion Index and Mean Index Adequacy (MIA).

All papers reviewed employ the same structure in the analysis of the data, as shown in Figure 12. The blue boxes indicate the actions that all papers undergo: data processing, method selection, clustering and validation of results. Some papers characterize the clusters (red box), enabling a description of the identified consumers, while others used the clusters thus obtained to segment new meters, followed by subsequent validation.

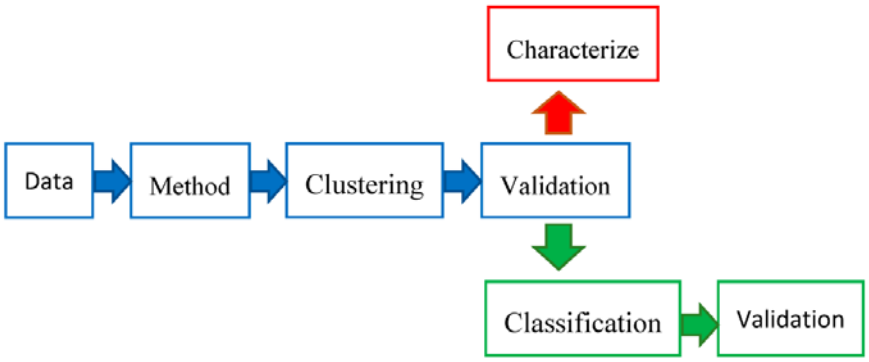


Figure 12 - Depiction of standard modelling structure. (Blue) Indicates the actions all papers go through (data-processing, method selection, clustering and validation). (Red) Some papers charachterize the identified clusters, usually applying external data. (Green) Applying the identified clusters to classify new consumption series to evaluate the applicability of the clusters. As presented in paper 1.

The review uncovered variations in the attention paid to systematic description of the data analyzed. This in turn restricts the transparency and reproducibility of the studies. Paper 1 presents a simple data description table allowing researchers to produce a compact data overview, acknowledging the preprocessing involved and its possible consequences. The data description was tested in papers 2 and 3 and presented therein.

4.1.4. Conclusion

The papers assessed all successfully produce consumption clusters through analyzes of smart-meter data. The clusters thus developed are of academic importance, but linking this to general applicability is largely lacking, partly due to two findings. First, the papers overlook the large overlaps between clusters, making it impossible to classify a new consumer uniquely. The clusters are represented by cluster means, while the cluster variance is largely neglected. Secondly, the stability of the identified clusters is seldom evaluated, nor are transitions of consumers between clusters over time.

Only one paper uses a methodology from time-series analysis, thereby acknowledging the time-series structure in smart-meter data. Paper 1 hypothesizes potential improvements to the current performance of consumption clustering by incorporating information about temporal aspects into the clustering.

One of the findings is the varying emphasis on data description, with few papers including detailed descriptions of data analyzed, while many papers produce inadequate descriptions enabling the reader to understand the data. This poses potential problems when trying to reproduce the study or evaluate the assumptions underlying model selection. The paper develops a data description table with thirteen

elements for authors to use when describing data to ensure that relevant information about smart-meter data is included in future papers.

4.2. Paper 2 - Electricity Consumption Clustering Using Smart Meter Data

This paper has been published in the MDPI journal *Energies*, an open-access journal.

4.2.1. Scientific Outline

The paper uses one week of smart-meter electricity data from over 32,000 households from the city of Esbjerg in southern Denmark to identify consumption patterns. The SydEnergi (SE) data analyzed are described in section 2.2 and in papers 2 and 4. The data selected for the clustering analysis in the paper are very homogeneous, only including households connected to the district heating system. The research is specifically aimed at investigating temporal dependencies in smart-meter electricity data – that is, the existence of autocorrelation – and possible ways to incorporate this in the clustering. The data used are presented in the data description table (Table 16).

DATA DESCRIPTION	VALUE
COUNTRY	Denmark
REGION	Region Syd (Region South) postal codes: 6700, 6705, 6710, 6715 (City of Esbjerg)
SUPPLIER	SydEnergi (SE) Electricity Utility
INITIAL SIZE	34,418 smart meters
CLEAR REDUCTION	See Table 2 in the paper.
MISSING VALUES	70 smart meters
FINAL SIZE	32,241 smart meters
RECORDING FREQUENCY	60 minutes
START	10th of January 2011
END	16th of January 2011
LENGTH	168 observations (hourly readings) per smart meter
TYPE	Single family houses (18,058 initial size) Apartments (15,721 initial size) heated through district heating.
REFERRAL	Data not referenced before.

Table 16 - Data description table of SydEnergi data utilized in paper 2 the table concept is introduced in paper 1. Thirteen distinct features of the data are shown in the table. The table creates an overview of the analyzed data. As presented in paper 2.

4.2.2. Methodology

Exploiting that the SE data includes one year of readings for each meter, making it possible to identify meters in the period analyzed that exhibit vacant consumption and to remove them from the analysis. This is achieved by identifying households that increased their consumption by more than 200% week-on-week, as shown in Figure 13.

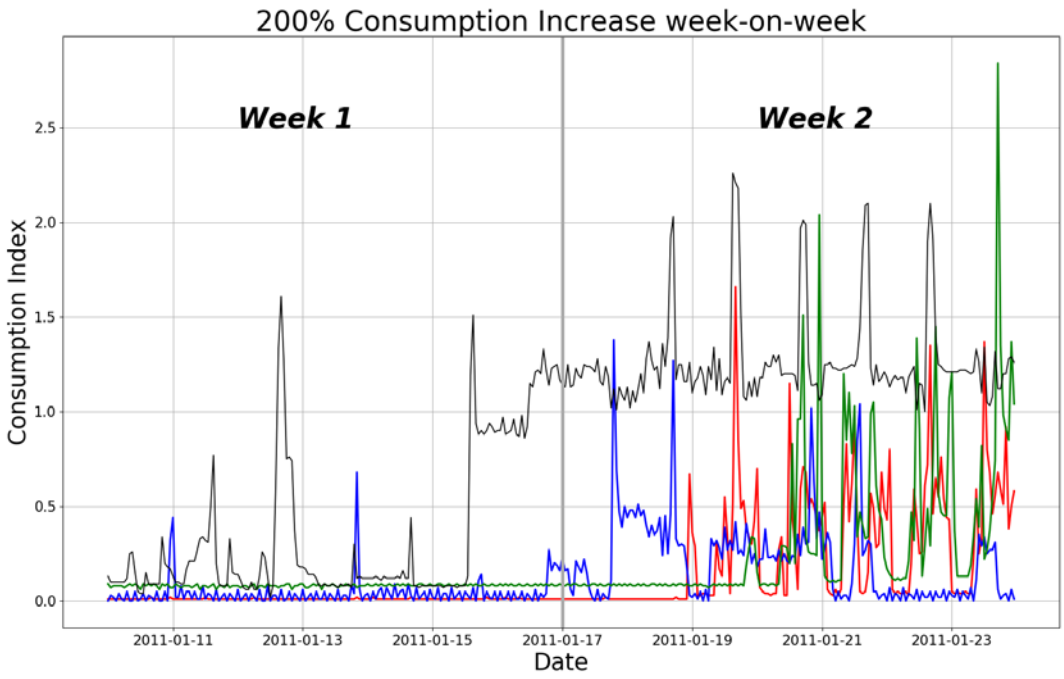


Figure 13 - Week-on-week consumption change for four meters, all demonstrating a 200% increase in consumption, indicating a vacant to non-vacant consumption pattern. As presented in paper 2.

Autocorrelation coefficients are calculated for each meter to establish the existence or non-existence of autocorrelation in smart-meter data. K-Means is selected as the clustering algorithm due to its prevalence in the consumption clustering literature. As described in section 3.2, this clustering method is unable to account for autocorrelation potentially residing in the smart-meter data. Methods applied to remove autocorrelation in the data include wavelet feature extraction and autocorrelation feature extraction. The clustering solutions adhere to the process chart outlined in Figure 14. All clustering solutions presented in the paper contain the steps described in the blue boxes: data processing, preparation and clustering, three different preprocessing methods of the input data are applied, autocorrelation features, normalization and wavelet features. Four CVIs are used to validate and select the clustering solutions: The Cluster Dispersion Index (CDI), the Davies-Bouldin Index (DBI), Mean Index Adequacy (MIA) and the Silhouette index.

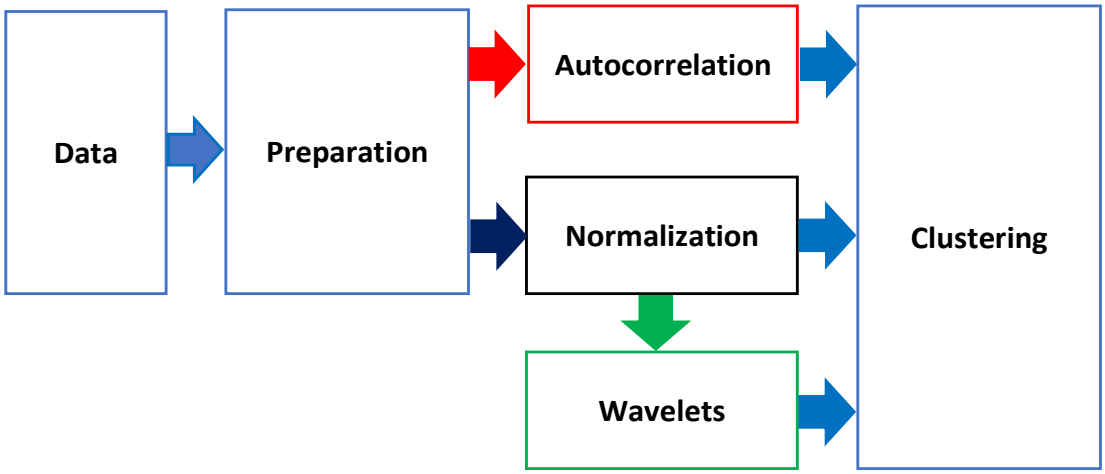


Figure 14 - Methodology flow chart. This chart illustrates the different data-processing methods used. The (blue) boxes indicate processes to which all methods were applied, namely data, preparation and clustering. After preparation, autocorrelation (red) indicates the extraction of autocorrelation features. Normalization (black) was applied both as a sole processing method, but also in preparation for wavelet transformation (green). As presented in paper 2.

4.2.3. Results

The paper demonstrates the existence of autocorrelation in smart-meter electricity data, also showing that different meters exhibit different degrees of autocorrelation. These differences indicate variations in the consumption-generating process and provide evidence of distinct consumption patterns. Figure 15 and Figure 16 show the consumption of individual meters during week two of January 2011. Going from left to right are the consumption measurements, the autocorrelation coefficients with confidence intervals and finally the retained significant autocorrelation coefficients. There are distinct differences in the figures and the autocorrelation.

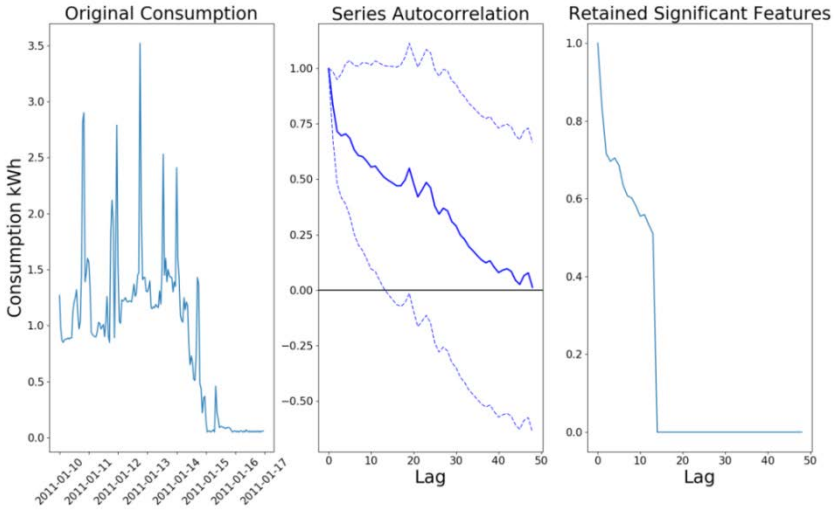


Figure 15 - (left) the original consumption profile, (middle) the solid line is the autocorrelation coefficient, dashed lines are the 95% confidence intervals, (right) significant autocorrelation coefficients retained as meter features and applied as input to K-Means clustering. The significant lags only include the first fourteen lags, indicating no recurrent pattern. As presented in paper 2.

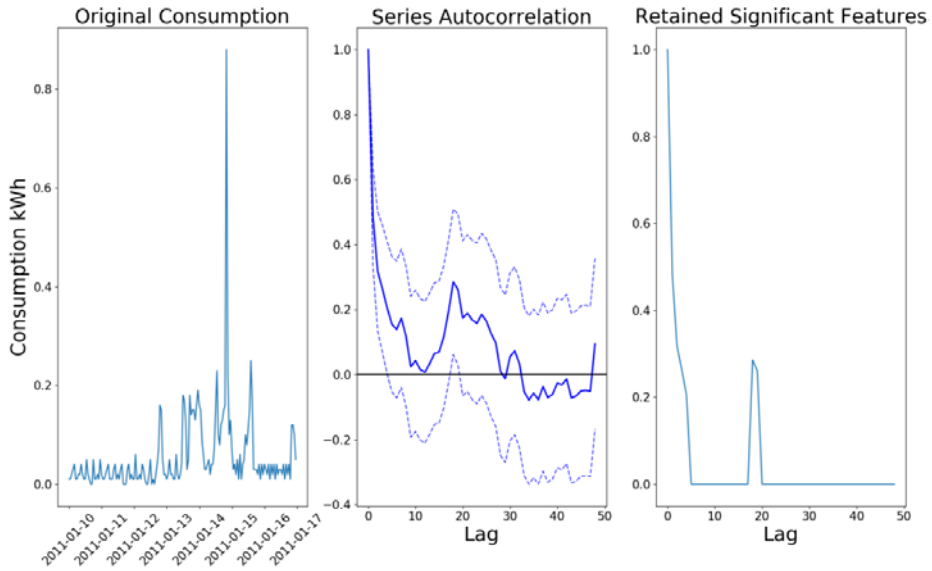


Figure 16 - (left) the original consumption profile, (middle) the solid line is the autocorrelation coefficient, dashed lines are the 95% confidence intervals, (right) significant autocorrelation coefficients retained as meter features and applied as input to K-Means clustering. The significant lags include lags from the first five lags and a recurrence at around lag 20, indicating some periodicity in the consumption. As presented in paper 2.

Having established the existence of temporal dependence by way of autocorrelation in the smart-meter electricity data, paper 1 found no evidence in the review of prior verification and no acknowledgement of time dependence through autocorrelation. Preprocessing of the data using the autocorrelation features extraction or wavelet features extraction before applying it as input to the K-Means clustering algorithm enables management of temporal dependence.

The wavelet feature method compresses the signal but retains the original structure of the consumption signal, thereby delivering clustering results that resemble results obtained through normalizing data. Figure 17 shows the CVI developments for normalized and wavelet transformed input data. Both exhibit fluctuating structures across the entire definition from two to 36 clusters, resulting in ambiguous cluster selection. The inability to cluster can be attributed to the homogeneous nature of the dataset, consisting of households in Esbjerg connected to the district heating system.

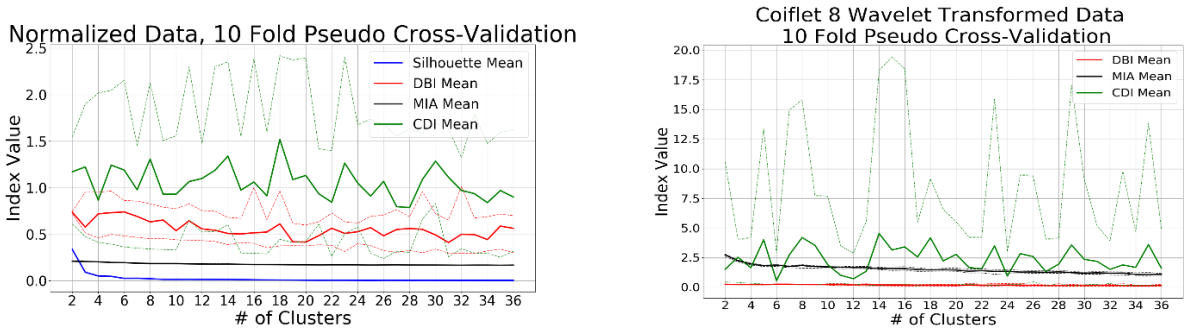


Figure 17 - (left) the CVI development as a function of clusters for normalized data. There is no apparent optimum; (right) CVI development as a function of clusters for wavelet transformed input data. Wavelets also indicate no apparent cluster optimum.

The autocorrelation features extract information about the underlying process that generates the consumption. By using the autocorrelation features as input to the K-Means clustering, the process type is clustered rather than the specific meter. Figure 18 shows the development of the CVI as a function of clusters. There is a very distinct “elbow” break, indicating that the optimum number of clusters is twelve. This is a considerable improvement compared to wavelet features and normalization, neither of which was able to identify the optimum number of clusters.

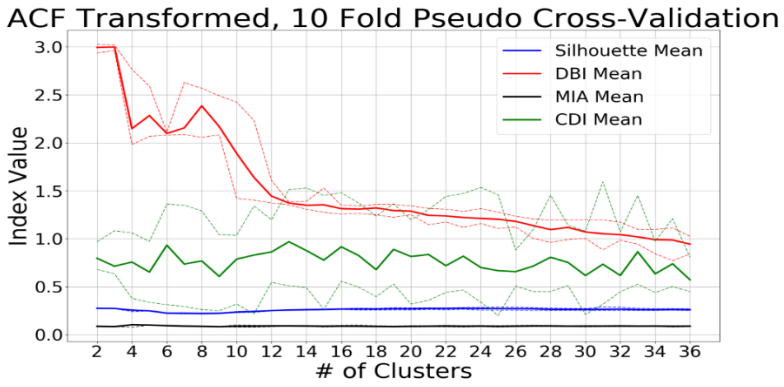


Figure 18 - Cluster validation index development for the autocorrelation features (ACF). The DBI index shows a distinct “elbow” break at twelve clusters. As presented in paper 2.

The resulting cluster means are shown in Figure 19. Ten of the twelve clusters reveal a similar structure, though different in respect of the impact of the five hours immediately past. A recurrent structure is observed around the 24-hour lag, showing daily repeatability in consumption patterns.

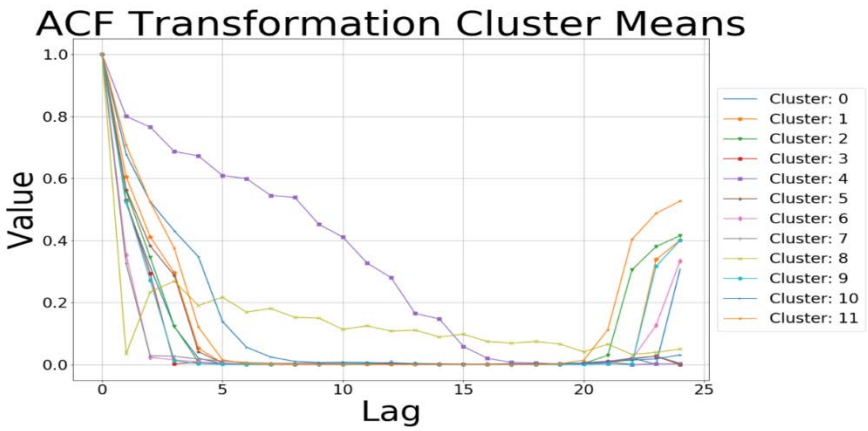


Figure 19 - Plot of the twelve-autocorrelation feature cluster means (ACF) identified using the CDI. Clusters 4 and 8 are distinctly different, showing linear decline and no recurrence. The remaining clusters exhibit a largely similar structure, with different values, and a different lag for recurrence. As presented in paper 2.

Investigating the composition of the twelve clusters reveals a balanced solution with ten large clusters and two smaller clusters. The clusters are similar in size and evenly distributed across both postal districts and dwelling type. The resulting composition of all twelve clusters can be examined in Table 17.

Cluster Composition			Dwelling Type		Postal Code in Esbjerg			
Cluster	Size	% of Total Data	Apartments	Houses	6700	6705	6710	6715
0	3198	9.92%	1244	1954	1396	571	754	477
1	2456	7.62%	851	1605	976	460	609	411
2	3342	10.37%	1240	2102	1427	603	798	514
3	3988	12.37%	1920	2068	1953	739	763	533
4	239	0.74%	117	122	127	36	45	31
5	4295	13.32%	1854	2441	1956	846	888	605
6	3014	9.35%	1616	1398	1522	586	489	417
7	3590	11.13%	2237	1353	1976	674	539	401
8	405	1.26%	300	105	256	63	46	40
9	3703	11.48%	1476	2227	1568	670	868	597
10	1794	5.56%	859	935	875	344	347	228
11	2217	6.88%	946	1271	940	462	488	327
Total	32,241	100.00%	14,660	17,581	14,972	6054	6634	4581

Table 17 - Cluster composition table of the twelve different electricity clusters. Only clusters 4 and 8 are markedly different from the rest, with a very small cluster size. The remaining cluster sizes are well-balanced across all parameters. As presented in paper 2.

4.2.4. Conclusion

The paper demonstrates the existence of autocorrelation in smart-meter electricity data, which has not been shown before [35], and proposes methods of successfully enabling the K-Means algorithm to account for autocorrelation through careful preprocessing of the input data. Preprocessing the data through autocorrelation features produces balanced clustering solutions with distinct clusters. The wavelet features produce solutions resembling normalized clustering, both being unable to produce unambiguous clusters. This can be attributed to the homogeneity of the data selected for clustering, but autocorrelation features are able to produce distinct clusters. Both autocorrelation and wavelet features significantly compress the data but retain the ability to cluster: the compression improves clustering speed and, in the case of autocorrelation features, also improves clustering ability.

4.3. Paper 3 - Clustering District Heat-Exchange Stations Using Smart-Meter Consumption Data

This paper is published in the Elsevier journal *Energy & Buildings*.

4.3.1. Scientific Outline

Paper 3 investigates consumption clustering using smart-meter district-heating data from heat-exchange stations (HX). Heat-exchange stations are equivalent to electricity transformation stations and connect the 120°C transmission grid to the 80°C distribution grid. The paper investigates the existence of temporal components in smart-meter district-heating data, identifiable as autocorrelation, and examines methods of improving clustering in order to manage autocorrelation. It uses research conducted and methodology successfully applied in electricity consumption clustering to district-heating consumption clustering. The data analyzed in paper 3 are provided by AffaldVarme Aarhus (AVA) and contain hourly consumption readings from heat-exchange stations from January 2017, ultimately including 49 stations, with 744 hourly recordings per meter for the whole of January. The data are aggregated into districts and are non-sensitive. An overview is given in Table 18.

DATA DESCRIPTION	VALUE
TYPE	Smart-meter readings from district heat-exchange stations, exchanging heat from transmission to distribution grid. Supplying smaller geographical areas of residential and industrial consumers with heat.
COUNTRY	Denmark
REGION	Municipality of Aarhus
SUPPLIER	AffaldVarme Aarhus (AVA)
INITIAL DATA SET SIZE	53 District Heating-Exchange Stations, with 744 readings each.
EXCLUSION OF DATA	Meter **#130** was removed, as it is a large company heat-exchange station serving only one customer.
MISSING VALUES	Meters: **#118A**, **#136J**, **#147 were** discarded due to missing data for the whole of January. Meters: **#111C**, **#119**, **#133**, **#134**, **#135**, **#136**, **#148** and **#151** represent erroneous readings. Imputation is described in paper 3.
FINAL DATA SET SIZE	49 district heat smart-meters with complete data with 744 readings each.
RECORDING FREQUENCY	Hourly (sixty-minute intervals)
START	01/01/2017
END	31/01/2017
LENGTH	744 recordings per meter. Hourly recording for the whole of January.
REFERRAL	Data not referenced before.

Table 18 - Data description table summary of the AffaldVarme Aarhus (AVA) Heat Exchange station consumption. An adaptation from paper 3.

4.3.2. Methodology

The paper applies K-Means clustering to the AVA data, clustering the entire month of January. The data are filtered and preprocessed to ensure analytical quality through the removal of meters with missing values and imputing series means onto outlying values. Figure 20 shows meters exhibiting outlier values (left) and the mean corrected data (right).

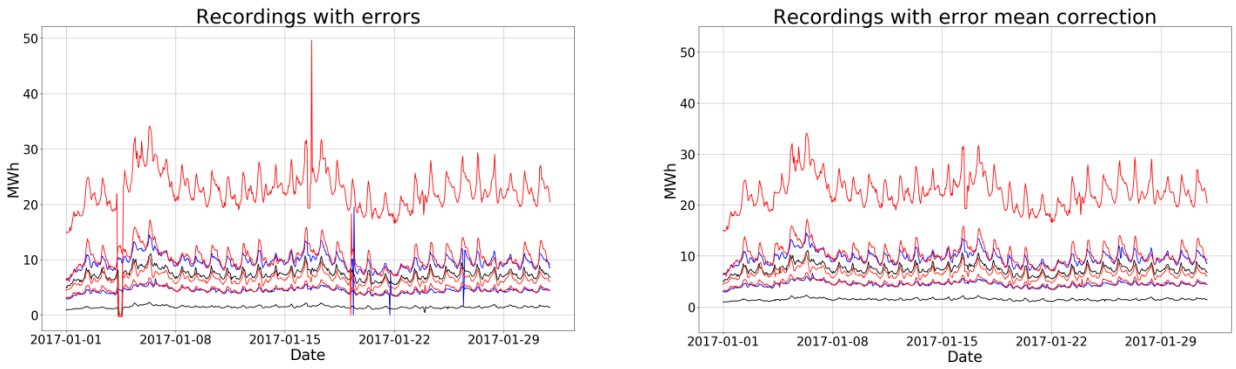


Figure 20 - Identification and imputation of outlier values: (left) the original raw data, (right) the series mean imputed data. The imputation is simple but visually leaves no undesirable artifacts. As presented in paper 3.

The data are prepared for analysis using four scaling methods: normalization, standardization, mean-centering and mean-divide. The clustering results are validated using four CVIs: MIA, CDI, DBI and Silhouette. A novel method of conducting unsupervised cross-validation is introduced using the CVI as

pseudo-response variables. The presence or not of a temporal component is investigated using autocorrelation. Autocorrelation feature extraction and wavelet transformation are applied to enable K-Means to handle autocorrelation in the data.

4.3.3. Results

The paper identifies the existence of autocorrelation in smart-meter district-heating data and successfully extracts significant autocorrelation features from the data as input for the K-Means clustering method. Figure 21 confirms the existence of autocorrelation and seasonality identified in heat-exchange station 145 Kolt.

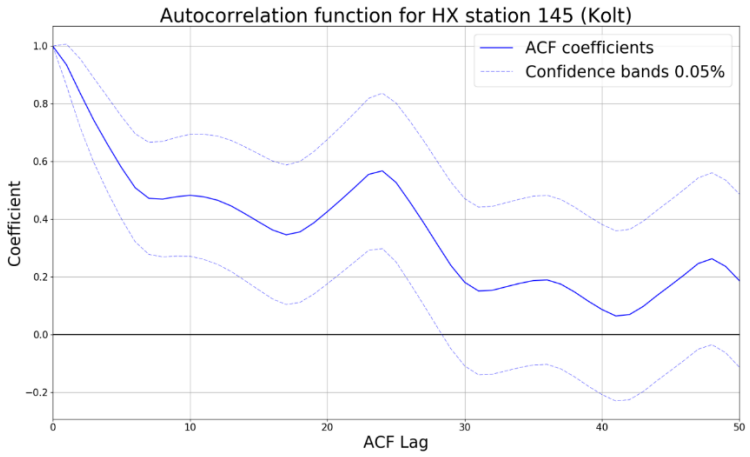


Figure 21 - Autocorrelation plot for heat exchange station 145 (Kolt) with fifty lags. Autocorrelation is indicated by the solid line, while the dashed lines indicate the 95% confidence intervals. Autocorrelation and seasonality components are visible in the figure. As presented in paper 3.

Clustering of consumption using normalized data results in a viable clustering solution, although the resulting clusters contain large variation such that the clusters overlap.

The wavelet transformation can account for the autocorrelation in the data and reduce the dimensions in the input data significantly from 744 readings to 161 coefficients. This reduces the computational cost of the K-Means clustering by several orders of magnitude. The clustering results from the wavelet transformation clustering are very similar to those from the normalized clustering. The similarity structure can be seen in Table 19, where almost all clusters are located in the diagonal.

METHOD		**WAVELET FEATURES**				
	Cluster #	0	1	2	3	Total
NORMALIZED	0	16			2	18
	1		4			4
	2	3		9		12
	3	1			14	15
	Total	20	4	9	16	49

Table 19 - Cluster overlap table between normalized and wavelet transformed data. Columns show wavelet transformed clustering, and rows show normalized clustering. Normalized and wavelet produce similar clusters with this data set. Nearly all normalized clusters are mapped 1:1 to the corresponding wavelet cluster. The two methods yield similar clustering in this case. An adaptation from paper 3.

Table 20 shows the autocorrelation features producing a different clustering solution that is more finely grained than that obtained from the normalized data. The normalized data suggest four clusters, while autocorrelation indicates seven. The difference in clustering is not just in the ability to sub-cluster the normalized clusters, as the method develops distinctly different clusters. The autocorrelation feature method transforms data and by design reduces the dimensions from 744 to 24, and with this data produces clusters that are distinctly different from the normalized clusters.

METHOD	AUTOCORRELATION FEATURES							
	Clusters #	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**Total**
NORMALIZED	**0**	7	3		1	3	2	2	18
	1				2		2		4
	2	2		2	1	4	3		12
	3	2	2	2	9				15
	Total	11	5	4	13	7	7	2	49

Table 20 - Cluster overlap table. The columns show the seven clusters from the autocorrelation feature (ACF) transformation clustering, while the rows show the normalized clustering. All normalized clusters are scattered across several ACF clusters, showing that the detailed ACF clustering is not just a subset of the normalized clustering but entirely different clusters. An adaptation from paper 3.

Though the autocorrelation feature clustering develops finer clusters than the normalized version, the paper is unable to show a statistical significance in their clustering performance between the three methods: normalization, wavelet transformation and autocorrelation features.

4.3.4. Conclusion

The paper confirms the existence of autocorrelation in smart-meter district-heating consumption data. This is an important finding, as it emphasizes the need for the selected clustering methods to incorporate temporal structures into the clustering to capture all the available information. Autocorrelation features and wavelet transformations are able to preprocess the data so that temporal dependencies are handled and included in the K-Means solution. Both methods reduce the dimensions and thus the computational effort, but the autocorrelation feature can visually deliver more distinct and more finely grained clusters.

4.4. Paper 4 – Stability of Electricity Smart Meter Consumption Clusters over Time

This paper has been submitted to the Elsevier journal Applied Energy.

4.4.1. Scientific Outline

Paper 4 evaluates the stability of electricity consumption clusters over time. This entails investigation of individual meters that are clustered together, regardless of the time of year. The data analyzed cover the whole of 2011 with hourly resolutions. To keep the computations feasible the year is divided into quarters: January -March, April-June, July-September and October-December. Each quarter is subdivided into weeks as the smallest time window in which to cluster, and each week contains hourly resolution as the smallest recording entity. Week 52 is excluded due to its radically different consumption pattern because of the Christmas holidays. The paper clusters and compares each week within each quarter to determine whether meters are clustered together throughout the quarter.

4.4.2. Methodology

The data analyzed in the paper are similar to those analyzed in paper 2, except that rather more data-cleaning was required to ensure all the included meters have readings across the entire year. The methodology for clustering likewise includes lessons from paper 2 in respect of preprocessing the data with the autocorrelation features prior to K-Means clustering and applying pseudo-cross-validation of the cluster validation indices in the selection of clusters. The paper develops and applies the Varatio methodology described in section 3.7 in assessing clustering stability.

4.4.3. Results

All four quarters exhibit differences in the weekly estimated optimum number of clusters. In general, however, six clusters appear as the majority across all four quarters. Within each of the four quarters the Varatio estimate for mapping is approximately 1:k (described in section 3.7), with a few mappings exceeding a 20% overlap. This indicates that the clusters are only valid for the week they are defined and are not generalizable to other weeks. This finding is recurrent throughout the four quarters of the year. Table 21 shows the Varatio estimates tabulated for three different weeks of quarter 2. The three tables show poor Varatio coefficients with all weeks in the same quarter.

Week 19 Varatio overlap with rest of weeks in Q2

Week	19 to 14	19 to 15	19 to 16	19 to 17	19 to 18	19 to 19	19 to 20	19 to 21	19 to 22	19 to 23	19 to 24	19 to 25	19 to 26
Cluster 0	8%	9%	8%	9%	9%	100%	8%	7%	9%	7%	10%	8%	11%
Cluster 1	6%	6%	7%	7%	6%	100%	8%	7%	7%	5%	8%	6%	9%
Cluster 2	14%	13%	7%	11%	13%	100%	13%	15%	10%	13%	12%	13%	12%
Cluster 3	17%	19%	16%	19%	22%	100%	19%	20%	19%	18%	18%	17%	19%
Cluster 4	4%	4%	4%	5%	4%	100%	3%	5%	5%	4%	6%	3%	4%
Cluster 5	9%	9%	8%	9%	10%	100%	9%	8%	9%	9%	10%	9%	10%

Week 20 Varatio overlap with rest of weeks in Q2

Week	20 to 14	20 to 15	20 to 16	20 to 17	20 to 18	20 to 19	20 to 20	20 to 21	20 to 22	20 to 23	20 to 24	20 to 25	20 to 26
Cluster 0	11%	11%	6%	9%	10%	11%	100%	12%	9%	11%	11%	12%	11%
Cluster 1	19%	19%	17%	20%	22%	23%	100%	20%	20%	18%	19%	18%	20%
Cluster 2	6%	6%	6%	7%	5%	7%	100%	6%	7%	5%	8%	6%	9%
Cluster 3	9%	10%	9%	10%	11%	10%	100%	8%	10%	10%	11%	9%	11%
Cluster 4	3%	2%	2%	5%	3%	3%	100%	5%	4%	2%	4%	2%	4%
Cluster 5	8%	9%	9%	9%	9%	9%	100%	8%	9%	7%	10%	8%	11%

Week 22 Varatio overlap with rest of weeks in Q2

Week	22 to 14	22 to 15	22 to 16	22 to 17	22 to 18	22 to 19	22 to 20	22 to 21	22 to 22	22 to 23	22 to 24	22 to 25	22 to 26
Cluster 0	6%	6%	6%	6%	5%	6%	7%	5%	100%	5%	8%	5%	8%
Cluster 1	10%	10%	10%	10%	11%	11%	11%	8%	100%	10%	11%	11%	11%
Cluster 2	16%	17%	15%	17%	20%	20%	17%	17%	100%	17%	18%	16%	18%
Cluster 3	8%	8%	8%	9%	8%	8%	8%	7%	100%	7%	10%	7%	11%
Cluster 4	4%	4%	3%	4%	3%	4%	4%	4%	100%	4%	7%	4%	5%
Cluster 5	12%	11%	8%	10%	10%	11%	13%	13%	100%	14%	13%	13%	12%

Table 21 - Varatio coefficients for each cluster combination in weeks 19, 20 and 22. Dark green indicates a 50%+ Varatio coefficient. Light green indicates Varatio estimated at between 20-50%, yellow at 10-20%, light red at 5-10% and dark red at <5% of maximum variance as defined by Varatio. In all three selected weeks of Q4, the mapping is approximately 1:k. The data is rounded to nearest integer value. As presented in paper 4.

4.4.4. Conclusion

Regardless of the quarter analyzed, the paper was unable to identify any weeks producing time-stable clusters. Each week's clusters are valid only for precisely that week. The mapping investigated in the paper unanimously suggests that the clusters are mapped approximately 1:k throughout the entire year. This finding implies that the clustering produced by applying the K-Means algorithm to smart-meter data produces non-viable clusters that are neither generalizable nor applicable in a practical setting for utilities.

4.5. General Paper Discussion

The papers in this thesis have been designed to maximize their contributions to the research objectives and deliver a general coherence of the study. Subsequent papers apply knowledge gained in previous papers thereby contributing to a natural progression throughout the study.

Paper 1 creates a comprehensive and coherent summary of the field of smart meter analysis, which allows for identification of gaps in the current methodology applied. Paper 2 applies the state of the art of the field to similar data from a Danish electric utility and tries to improve the methodology by accounting for temporal structures. Paper 3 extends the knowledge from electricity consumption data to district heating consumption data and shows similarities between the data types. Finally, paper 4 investigates the stability of the cluster solutions. Throughout all the papers, methods have been modified or developed to narrow or close identified gaps. The overall progression of the papers is shown in Figure 22.

Figure 22 - Paper progression: Paper 1 generates an overview of the state of the art in smart meter data analysis. Paper 2 applies the knowledge to Danish consumption data. Paper 3 Investigates if the methododology is applicable to district heating. Finally paper 4 studies the time persistance of cluster solutions.

Paper 1 systematically identifies important papers; it does so by applying a modification of Okoli's method for systematic literature reviews. This way of conducting a review was chosen as it presents a clear procedure for identifying studies and evaluating the relevance. Different strategies exist for literature reviews, but this method delivers repeatability, and a transparency to the process. It allows for later amendments either by updating the review with published material or widening through inclusion of new key-phrases. The process is largely deterministic, though the assessment of paper relevance is ultimately at the author's discretion. This bias is affecting the study, through keyword selection, inclusion criteria, and protocol, which directs the focus of the study. To ensure repeatability all datasets collected from web of science have been kept for later reference, and each step of the paper review processing has been documented and archived and it has often been revisited to investigate why some papers was not included in the review.

Where paper 1's target is a comprehensive coherent summary of the field, paper 2 and 3 are focused on the applicability of the methods to the major component of the Danish energy system; electricity and district heating.

Paper 2 analyzed subsamples of electricity consumers connected to district heating and living in downtown Esbjerg. This subset is very homogeneous compared to the overall dataset but still expected to include many different consumer types. Clearly a more inhomogeneous subset of the data could have been beneficial in achieving more successful clustering. The current subset selected represents a large portion of household in Denmark, and applicable clustering solution of this subset is important in identification of flexibility. It is important to develop methods that can produce distinct clusters on this group of consumers due to their large prevalence in the Danish population.

Not having encountered district heating in the review of paper 1, the analysis of this type of data is novel and important. It is not evident that methods applied for fast moving energy as electricity are applicable to the much slower district heating. Mimicking the analytical process from electricity consumption clustering to district heating ensures a natural extension to the literature of electricity smart meter clustering to district heating, while at the same time uncovering if these data types can be analyzed using comparable methods.

The district heating data analyzed are aggregated but still represent an energy type not encountered in the field before. Data acquisition of household level smart meter consumption was unmanageable within the time frame of the PhD, due to compliance with data privacy regulation. The current aggregated level of clustering contributes to the overall literature, but household level detail could provide even more information and detailed analysis. The aggregated consumption enables clustering the differences between districts rather than individual consumers.

Few studies identified in paper 1, evaluate the resulting cluster solution or apply it to cluster new meters. Paper 4 develops and successfully applies a methodology for evaluating if clusters are persistent in time. For the SE electricity data the results are discouraging, with no cluster evaluated being persistent. The Varatio quantifies what paper 1, 2 and 3 have been discussing about performance of K-Means clustering of smart meter data. It does not produce results which are entirely unexpected, but the degree to which Varatio show scattering of clusters across time is surprising. Varatio needs more testing on other smart meter consumption data to see if the findings are general across energy types, and to evaluate the method.

Having been unsuccessful in the review to identify papers that are able to create truly distinct clusters, paper 2, 3 and 4 tries different methods to create unique clusters and assess the resulting cluster stability. The thesis identifies gaps and provides methodology for improving the clustering and assessing the stability of the clusters. The thesis is unable to develop or identify current methods which produce distinct clusters. This and other studies show there is plenty of structure in the data, but the current methods are not able to exploit this information and produce distinct clusters.

5. Discussion

In today's society, data is perceived as "The New Gold" [62], [63], holding out the promise of disrupting well-known societal structures and revolutionizing industries [64]. Smart-meter data have been hailed as the new premier product for electric utilities [21], [23], [22] relegating electricity to a mere commodity needed to obtain valuable high frequency consumption data.

Current legal requirements to introduce smart-meters and store consumption data have increased cost and operational complexity for the individual utilities without enabling clear potentials for utilities to generate imminent income from the data. This thesis has analyzed smart-meter electricity and smart-meter district heating consumption data to evaluate the potential for consumption clustering and highlight pitfalls and limitations. Diverse tariffs catering for different profiles could incentivize consumer flexibility, thereby optimizing grid operation and reduce maintenance and operational costs for the utility.

Renewable energy sources, especially wind energy, introduce volatility into the electricity grid, smart meter data can be used to create tariff structures that incentivize flexible demand, to stimulate consumption during high production times. Electrification of the transportation sector will only increase the need for flexibility. Different tariff schemes able to regulate consumption are needed for the future electricity grid to sustain the increased demand when electrification of transportation is realized. Tariffs as incentives for flexible consumption can help reduce electricity grid strain and save end-users of large costs related to strengthening the grid. It is expensive and infeasible for utility companies to develop tariffs for individual household to leverage consumption flexibility. Smart-meter electricity consumption data potentially enables utilities to identify similar consumption patterns, making it possible to identify consumption patterns and target similar consumers with a relevant tariff. Development of these tariffs requires consumption insights which smart-meters can provide.

Flexibility in the energy system does exist [65], and the individual elements of the energy system are highly optimized. The individual households have little influence on the overall system. Smart meter data can make individual households an integral part of the overall system and allow for consumption flexibility rather than just system flexibility. Legislation must allow for the introduction of technological solutions and enable frameworks for creating incentives; in this regard smart-meters are a tool for creating and enforcing tariff strategies and not the solution itself.

There are differences in energy consumption between households, even when no electricity is used for heating. Papers 2 and 3 show this in a Danish context. While paper 1 identifies studies that successfully identify electricity consumption clusters, the clustering solutions are unable to produce clusters that are valid outside the period analyzed. Smart-meters enable unprecedented detail about consumption, data that have the potential to provide deep insights into consumption patterns and can be expected to help identify flexibility. Analyzing the current state of the art in smart meter consumption clustering, this thesis shows that creating stable clusters with the currently prevailing methodology from smart-meter consumption data is not a trivial task.

The flexibility solution identified for household consumption cannot be allowed to degrade the comfort levels of inhabitants. The scheme for governing this flexibility must be fair and balanced; otherwise consumers will reject and counteract the scheme due to egocentric behavior. This has been observed in a recent study where a small village consumption flexibility was analyzed, the imposed tariff structure was unbalanced and resulted in inflexible and shortsighted consumption behavior [66].

Even though Danish consumers pay some of the lowest electricity prices in Europe [67], the current flat taxation (including VAT) of 70% [68] of electricity in Denmark amounts to the highest net charge in the same region [69]. The remainder of the electricity bill is split evenly between the DSO and the supplier. The margins for the utilities are low [2], with little incentive for consumers to switch supplier. There is little room for either to introduce incentive for consumers to behave flexibly. E.g. a household is offered reduced prices in selected time slots resulting in shifting 20% of consumption to this time slot. This change only affects the 15% of the overall electricity bill; consequently, the household only gains as little as 3% overall reduction on the electricity bill due to fixed costs. This is because of high fixed costs makes the first kWh electricity MWh heat expensive, especially for small households [70]. Coupled with the relative small impact on household's financial situation, a 3% saving is comparably small incentive compared to reduced convenience and increased micromanagement. To realize the potential of smart-meters consumption data, the electricity taxation must enable incentive structures for consumers to behave flexibly. Smart-meters alone can only identify when energy is consumed and create the insights needed for developing tariff structures that motivate consumption flexibility. If incentives are not enabled, smart-meters will remain nothing more than an advanced system of measuring consumption.

The current methods applied to consumption clustering using smart-meter data do not leverage the time series structure of the data. Few methods exist for clustering time series [71], and further research into time series clustering is needed such that the clustering will include the intrinsic information like autocorrelation structures. This thesis has proposed several methods for improving clustering results by introducing preprocessing of the input data, thus enabling K-Means to include temporal information. Paper 2 and 3 used preprocessing of data to improve the definition of the resulting clusters; both papers improved the clustering compared to non-preprocessed solution. The papers 2 and 3 were unable to show that the reduction in variance achieved by preprocessing the data enabled statistically improved clustering, meaning that the variance, though reduced, still produces overlap between clusters. Regardless of the proposed preprocessing; K-Means is capable of clustering smart meter consumption data, but is unable to create distinct clusters, nor does it include important inherent information from the data.

Though the K-Means can produce clustering solutions based on smart-meter data, the resulting solution is so far of mere academic importance, illustrating the robustness and clustering ability of the method. However, the practical applicability of the clustering is unclear, as the within-cluster variance is substantial, and large enough that clusters overlap statistically. Consequently, the overlap between clusters results in a lack of discriminatory power, making it difficult to create viable consumption clusters. All four papers engage in a discussion on how to reduce the within-cluster variance and thus improve the identifiability of the individual clusters but without reaching conclusive results. Papers 2, 3, and 4 all introduce methodology for either improving clustering or evaluating cluster stability, but the methods introduced are unable to supply statistically improved results compared to previous research. Further research into methods able to create distinct smart–meter consumption clusters and statistical tests for testing similarity between time-series are needed.

Paper 4 introduces the Varatio method for evaluating cluster stability, showing that the clusters created using K-Means are only valid for the time selected for the clustering. This result can be surprising factoring in the decade of successful clustering using K-Means and electricity smart-meter data. Some papers have inadvertently proposed this result through the visual inspection of clusters hinting to overlapping and unambiguous clusters [28] though successful in producing clusters their stability was never investigated.

Few papers have had access to datasets which enable the analysis of cluster stability; this has propelled the successful clustering of smart-meter data but without the ability to prove the stability outside the realm of

the cluster validation indices and the brief period analyzed. The indices only supply information about the optimum clustering given the data at hand, and do not provide information about cluster overlap nor stability over time. The recognition that overlapping clusters and optimal number of clusters are not mutual exclusive events is vital in the four papers included in this thesis. It shows that cluster validation indices can hint to the optimal number of clusters without necessarily enabling unique clusters.

Smart-meters record consumption at very high frequency, but there is no consensus on recording frequency for consumption clustering. Popular choices are fifteen and sixty minutes, but no research identified has investigated the impact of the frequency on clustering capability. The currently selected methods for clustering smart-meter consumption are readily available in most analytical software, but as the papers 2, 3 and 4 shows the methods are producing clusters with limited applicability. There is a need for theory and tools for comparing realizations of time series thus quantifying differences. Missing is also a statistical framework allowing for statistical comparison of clusters of time series such that clustering methods can be compared and resulting clusters can be evaluated.

K-Means is simple to apply which is possibly the reason for its widespread application in smart-meter consumption clustering, but its ability to identify clusters is not equivalent to the creation of distinct and unambiguous clustering results. This thesis recommends that future smart-meter consumption clustering develops and employs other algorithms than K-Means as the solutions created with K-Means are often not generalizable, and the clusters produced may not be valid in a larger context. The problem of the poor clustering amounts to the lack of methods able to cluster time series effectively. There is a large body of research into the analysis of time series, enabling identification of models and forecasting of event, but within time series there is a gap in the literature regarding the evaluation of (dis)similarity between time series. Without dedicated methods for clustering time series; K-Means and business rules generated from customer knowledge are potentially the best options currently available for clustering smart-meter consumption data with all the pitfall and unambiguity this entails.

6. Conclusion and Outlook

This thesis has via five research objectives analyzed the applicability of smart-meter electricity and district heating smart-meter consumption data for clustering. It has shown that Danish smart-meter electricity consumption data behave equivalently to previously published research papers and that the data does demonstrate evidence of consumption profiles.

Paper 1 of this thesis contributes to the literature by evaluating the current state-of-the-art in smart-meter consumption clustering. The evaluation is conducted by applying Okoli's method for a systematic literature review. Paper 1 has generated an overview of the main methodologies applied in smart-meter electricity consumption clustering, with simple methods like K-Means and Hierarchical clustering as the most prevalent. Neither is designed for clustering time series data such as smart-meter data. Furthermore, paper 1 identified many cluster validation indices for estimating the optimum number of clusters when applying unsupervised clustering. The main conclusion from the review in paper 1 is the capability of K-Means to successfully calculate clusters from smart-meter data, but that current research papers are not harnessing intrinsic information in the data, such as autocorrelation to improve the clustering.

The lessons from the review are used in paper 2 to cluster Danish smart-meter electricity consumption data. The K-Means algorithm is used for clustering to ensure comparability to current studies and the clustering solutions are equivalent. No papers in the review investigate the smart-meter data for autocorrelation, making paper 2 the first paper to show potential for autocorrelation in smart-meter electricity data. Paper 2 applies the four most prevalent cluster validation indices and introduces autocorrelation features and wavelet features to enable K-Means to manage autocorrelation in the clustering. The resulting clustering is finer grained than clustering omitting the autocorrelation.

Paper 3 is reproducing paper 2 with smart-meter district heating consumption data. The same cluster validation indices and clustering methodology is applied. This paper also proves the existence of autocorrelation in smart-meter district heating data. As with electricity consumption data the clustering of district heating data is improved by applying autocorrelation features and wavelet features. Paper 3 shows that the methodology applied for electricity data consumption clustering is readily applicable for district heating smart-meter data.

The papers 1, 2 and 3 all question the generalizability and applicability of the clusters created. Paper 4 develops a method; Varatio which can evaluate cluster stability over time. The paper shows how the clustering of one week of electricity data does not generalize to other weeks of the same year. The clustering solutions generated via K-Means are difficult to generalize, meaning new smart-meter consumption data cannot easily be categorized according to some previous clustering solution generated by K-Means.

Smart-meter electricity and district heating consumption data collected for billing purposes does exhibit different consumption patterns, however the prevalent methods currently applied in smart-meter consumption clustering struggle to produce viable clustering solutions which generalize across time and data. This thesis has not been able to identify clustering algorithms capable of producing unambiguous clusters from smart-meter consumption data. The currently applied methods do not produce sufficiently distinct clusters for the solutions to be feasible in a practical context retaining smart-meter consumption clustering as an academic exercise at the moment.

Outlook

Through the four papers this thesis has shown that there is a need for further research into smart-meter consumption clustering methods. The current methods applied are able to produce distinct clusters, but with large variation making the attainable clusters academically relevant, but their practical applicability questionable. Paper 1 indicates several unanswered questions regarding smart-meter consumption clustering, relating to time interval selected, distinguishability of clusters, methods for testing differences etc.

Although this thesis, using Danish smart meter data, presents results which are in line with the current literature, the resulting cluster solutions using state of the art methodology have little practical applicability. The methodology for clustering time series data must improve the handling of variation to enable better distinction between clusters. Papers 2, 3, and 4 shows there is much information hidden in the smart meter consumption data, but this information is for the moment not applicable.

As smart meters produce a data type encountered in many fields; energy consumption, log-files, finance, computer security etc. the problem of distinct clusters is general, and many fields are studying it. I have no doubt that incremental improvement to unsupervised clustering, such as this thesis; will eventually enable creation of distinct clusters with this type of data, allowing for applicability outside academia.

So, if the current clusters are non-stable and not distinguishable where does it leave analysis of smart meter consumption data? Increasing the discriminatory power by including e.g. socio-economic and demographic data can potentially improve the applicability of the clustering. The coupling of different types of energy consumption data is also an interesting prospect. At individual household level smart meters deliver a data structure that has been extensively studied in other fields. Methods from time series allow for forecasting of individual households, while the field of statistical quality control delivers tools for detecting shift in consumption patterns. At cluster level the methods are not yet able to deliver results that are comparable to what can be achieved for the individual meter.

Another scenario is moving the clustering of consumption from the utilities to 3rd party vendors who can deliver tailor-made services that cater to specific types of customers. Danish electricity smart meter legislation allows for such constellations. The services can be available as opt-in services where the consumer supply consumption data, to the 3rd party which for a fee delivers a service to the household. This approach disregards the aim of the thesis of identifying clusters, but enables application of the meter data for the benefit of individual households. It could potentially provide security services to households of elderly or disabled people who supply data for profiling such that event not conforming to their usual schedule flags warning at the service provider, who can then call for assistance. Other potential application is the possibility of reduction in spending on energy through energy improvements. These improvements can be identified from consumption data coupled with information about the house and personal data which the consumer delivers.

Finally, the smart-meters harvest a resource which in today's society is sought after and debated, personal data. Possibly the meters are able to help identify if elderly discourse from their daily routine and might be in need of assistance. Conversely, smart-meters offer the threat of external surveillance of household's consumption data for misuse. Data centers as the Datahub from energy will become increasingly interesting as target for attack, further increasing the operational cost and risks. Is the perceived intrusion on personal privacy by deep knowledge of individual electricity consumption from the smart-metering,

surpassed by the societal benefits? A discussion about the limits to the application of smart-meter data is needed.

References

[1] D. Government, “Energy Strategy 2050 - from coal, oil and gas to green energy,” no. February 2011, 2011.

[2] Dansk Energi (Danish Energy Association), “Giv energien videre - nye energipolitiske visioner og udfordringer 2020 - 2030,” Copenhagen, Denmark, 2015.

[3] Central Statistics Office, *Energy statistics 2016*. 2016.

[4] T. J. Stine Jacobsen, “In windy Denmark, clouds clearing for solar power,” *Reuters*, 2017. [Online]. Available: https://www.reuters.com/article/us-denmark-renewables/in-windy-denmark-clouds-clearing-for-solar-power-idUSKCN1C220X.

[5] Euronews, “Can Denmark’s solar power solution be a blueprint for sun-starved countries?,” *Euronews*, 2017. [Online]. Available: http://www.euronews.com/2017/04/27/denmark-is-using-solar-to-heat-and-power-all-in-one-plant. [Accessed: 07-Mar-2018].

[6] A. Neslen, “Wind power generates 140% of Denmark’s electricity demand,” *The Guardian*, 2015. [Online]. Available: https://www.theguardian.com/environment/2015/jul/10/denmark-wind-windfarm-power-exceed-electricity-demand. [Accessed: 18-Jun-2018].

[7] Dansk Energi and Energinet.dk, “Smart grid i Danmark 2.0,” 2013.

[8] D. Fjernvarme, “Fakta om fjernvarme,” 2017. [Online]. Available: http://www.danskfjernvarme.dk/presse/fakta-om-fjernvarme. [Accessed: 16-Mar-2018].

[9] PlanEnergi, “Long term storage and solar district heating,” 2016.

[10] Ramboll, “World largest thermal heat storage pit in Vojens,” *State of Green*, 2015. [Online]. Available: https://stateofgreen.com/en/partners/ramboll/solutions/world-largest-thermal-pit-storage-in-vojens/. [Accessed: 16-Jul-2018].

[11] “IPower.” [Online]. Available: https://ipower-net.weebly.com/. [Accessed: 13-Jun-2018].

[12] “Ensymora.” [Online]. Available: http://www.ensymora.dk/. [Accessed: 13-Jun-2018].

[13] “EcogridEU.” [Online]. Available: http://www.eu-ecogrid.net/. [Accessed: 13-Jun-2018].

[14] “Flexpower.” [Online]. Available: http://www.ea-energianalyse.dk/projects-danish/1027_flexpower_markedsdesign.html. [Accessed: 13-Jun-2018].

[15] “4DH.” [Online]. Available: http://www.4dh.eu/. [Accessed: 13-Jun-2018].

[16] “CITIES - Centre for IT-Intelligent Energy Systems in cities.” [Online]. Available: http://smart-cities-centre.org/. [Accessed: 18-Jun-2018].

[17] I. Larsen, “Bekendtgørelse om fjernaflæste elmålere og måling af elektricitet i slutforbruget,” *Energistyrelsen*, 2013. [Online]. Available: https://www.retsinformation.dk/eli/lta/2013/1358. [Accessed: 07-May-2018].

[18] Comissão Europeia, “Smart Metering deployment in the European Union,” *European Commission*, 2015. [Online]. Available: http://ses.jrc.ec.europa.eu/smart-metering-deployment-european-union. [Accessed: 18-Jun-2018].

[19] . EU Commission, “Smart grids and meters - European Commission,” *European Commission*, 2014. [Online]. Available: https://ec.europa.eu/energy/en/topics/markets-and-consumers/smart-grids-and-meters. [Accessed: 20-Dec-2016].

[20] SEAS-NVE, “SEAS-NVE Watts.” [Online]. Available: https://watts.seas-nve.dk/?lang=en. [Accessed: 18-Jun-2018].

[21] C. Douris, “Utilities Should Sell Customer Electricity Data,” *Lexington Institute*, 2017. [Online]. Available: http://www.lexingtoninstitute.org/utilities-sell-customer-electricity-data/. [Accessed: 13-Jul-2018].

[22] J. Mazurek, “The data treasure chest: Is there a market to sell utility data?,” *Accenture Blog*, 2016. [Online]. Available: https://www.accenture.com/us-en/blogs/blogs-utility-data-treasure-chest-there-market-sell-utility-data. [Accessed: 13-Jul-2018].

[23] J. Worland, “Your Utility Company Wants to Sell You More than Just Electricity,” *TIME*, 2016. [Online]. Available: http://time.com/4312285/utility-company-electricity-solar-power/. [Accessed: 13-Jul-2018].

[24] S. Z. Christophe Guille, “How Utilities Are Deploying Data Analytics Now,” *Bain & Company*, 2016. [Online]. Available: http://www.bain.com/publications/articles/how-utilities-are-deploying-data-analytics-now.aspx. [Accessed: 13-Jul-2018].

[25] G. Chicco, R. Napoli, and F. Piglione, “Comparisons Among Clustering Techniques for Electricity Customer Classification,” *IEEE Trans. Power Syst.*, vol. 21, no. 2, pp. 1–7, 2006.

[26] G. Chicco, R. Napoli, F. Piglione, P. Postolache, M. Scutariu, and C. Toader, “Load pattern-based classification of electricity customers,” *IEEE Trans. Power Syst.*, vol. 19, no. 2, pp. 1232–1239, 2004.

[27] F. Mcloughlin, A. Duffy, and M. Conlon, “Characterising domestic electricity consumption patterns by dwelling and occupant socio-economic variables : An Irish case study,” *Energy Build.*, vol. 48, no. July 2009, pp. 240–248, 2012.

[28] J. L. Viegas, S. M. Vieira, R. Melício, V. M. F. Mendes, and J. M. C. Sousa, “Classification of new electricity customers based on surveys and smart metering data Commission for Energy Regulation,” *Energy*, vol. 107, pp. 804–817, 2016.

[29] A. Ozawa, R. Furusato, and Y. Yoshida, “Determining the relationship between a household ’ s lifestyle and its electricity consumption in Japan by analyzing measured electric load profiles,” *Energy Build.*, vol. 119, pp. 200–210, 2016.

[30] M. Piao, H. Shon, J. Lee, and K. Ryu, “Subspace Projection Method Based Clustering Analysis in Load Profiling,” *Ieeexplore.Ieee.Org*, vol. 29, no. 6, pp. 2628–2635, 2014.

[31] J. Kang and J. Lee, “Electricity Customer Clustering Following Experts’ Principle for Demand Response Applications,” *Energies*, vol. 8, pp. 12242–12265, 2015.

[32] A. Kavousian, R. Rajagopal, and M. Fischer, “Determinants of residential electricity consumption : Using smart meter data to examine the effect of climate , building characteristics , appliance stock , and occupants ’ behavior,” *Energy*, vol. 55, pp. 184–194, 2013.

[33] J. D. Rhodes, W. J. Cole, C. R. Upshaw, T. F. Edgar, and M. E. Webber, “Clustering analysis of residential electricity demand profiles,” *Appl. Energy*, vol. 135, pp. 461–471, 2014.

[34] G. Coke and M. Tsao, “Random effects mixture models for clustering electrical load series,” *J. Time Ser. Anal.*, vol. 31, no. 6, pp. 451–464, 2010.

[35] M. Tureczek. Alexander and S. Nielsen. Per, “Structured Literature Review of Electricity Consumption Classification Using Smart Meter Data,” *Energies*, vol. 10, no. 5, p. 584, 2017.

[36] “Python 3.6.x.” 2018.

[37] W. McKinney, “Data Structures for Statistical Computing in Python,” *Proc. 9th Python Sci. Conf.*, vol. 1697900, no. Scipy, pp. 51–56, 2010.

[38] T. E. Oliphant, “Guide to NumPy,” *Methods*, vol. 1, p. 378, 2010.

[39] S. Van Der Walt, S. C. Colbert, and G. Varoquaux, “The NumPy array: A structure for efficient numerical computation,” *Comput. Sci. Eng.*, vol. 13, no. 2, pp. 22–30, 2011.

[40] L. G, W. F, G. R, W. K, O. A, and N. H, “PyWavelets - Wavelet Transforms in Python,” https://github.com/PyWavelets/pywt, 2006.

[41] S. Seabold and J. Perktold, “Statsmodels: econometric and statistical modeling with Python,” *Proc. 9th Python Sci. Conf.*, no. Scipy, pp. 57–61, 2010.

[42] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, “Scikit-learn: Machine Learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2012.

[43] E. Jones, T. Oliphant, and P. Peterson, “SciPy: Open Source Scientific Tools for Python.” 2001.

[44] J. D. Hunter, “Matplotlib: A 2D graphics environment,” *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 99–104, 2007.

[45] J. Lattin, J. D. Carrol, and P. E. Green, *Analyzing Multivariate Data*, 1. st., vol. 46, no. 2. Duxbury, 2004.

[46] J. Friedman and T. Hastie, *The Elements of Statistical Learning*, 1st ed. Springer, 2008.

[47] L. A. Barford, R. S. Fazzio, and D. R. Smith, “An introduction to wavelets,” *Hewlett-Packard Labs, Bristol, UK, Tech. Rep. HPL-92-124*, vol. 2, pp. 1–29, 1992.

[48] T. Cormen, R. Rivest, and C. Leiserson, *Introduction to Algorithms*, Second edi. The MIT Press, 2001.

[49] D. Arthur and S. Vassilvitskii, “How slow is the k-means method?,” *Proc. twenty-second Annu. Symp. Comput. Geom. - SCG ’06*, p. 144, 2006.

[50] J. J. López, J. A. Aguado, F. Martín, F. Muñoz, A. Rodríguez, and J. E. Ruiz, “Hopfield-K-Means clustering algorithm: A proposal for the segmentation of electricity customers,” *Electr. Power Syst. Res.*, vol. 81, no. 2, pp. 716–724, 2011.

[51] S. Park, S. Ryu, Y. Choi, J. Kim, and H. Kim, “Data-Driven Baseline Estimation of Residential Buildings for Demand Response,” *Energies*, vol. 8, pp. 10239–10259, 2015.

[52] P. O. Perry, “Cross-Validation for Unsupervised Learning,” Stanford University, 2009.

[53] D. c. Montgomery, *Statistical Quality Control*, 5e ed. Wiley, 2005.

[54] F. Morchen, “Time series feature extraction for data mining using DWT and DFT,” pp. 1–31, 2003.

[55] B. Hurwitz, “Introduction to Wavelets -part 2,” vol. 1989, pp. 1–9, 2000.

[56] G. P. Nason, *Wavelet Methods in Statistics with R*. 2008.

[57] L. Wasserman, *All of Nonparametric Statistics*, no. 1. Springer, New York, NY, 2006.

[58] C. Okoli and K. Schabram, "A Guide to Conducting a Systematic Literature Review of Information Systems Research," *Work. Pap. Inf. Syst.*, vol. 10, no. 26, pp. 1–51, 2010.

[59] Thomson Reuters, "Web of Science 1 Billion Cited References and Counting," *Thomson Reuters*, 2017. [Online]. Available: http://stateofinnovation.thomsonreuters.com/web-of-science-1-billion-cited-references-and-counting?utm_source=false&utm_medium=false&utm_campaign=false. [Accessed: 05-Jan-2017].

[60] A. Tureczek, "Clustering District Heat Exchange Stations Using Smart Meter Consumption Data," in *3rd International Conference on Smart Meter Energy Systems and 4th Generation District Heating*, 2017, p. 24.

[61] P. Gianniou, X. Liu, A. Heller, P. S. Nielsen, and C. Rode, "Clustering-based analysis for residential district heating data," *Energy Convers. Manag.*, vol. 165, no. December 2017, pp. 840–850, 2018.

[62] J. M. Ron Bergers, "The New Gold," *Deloitte*, 2017. [Online]. Available: https://www2.deloitte.com/nl/nl/pages/data-analytics/articles/the-new-gold.html#. [Accessed: 03-Dec-2018].

[63] Leader, "The world's most valuable resource is no longer oil, but data," *Economist*, 2017. [Online]. Available: https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data. [Accessed: 13-Jul-2018].

[64] DAmon Lapping, "How Big Data Analytics is Disrupting the Energy Industry," *Disruptor Daily*, 2018. [Online]. Available: https://www.disruptordaily.com/big-data-analytics-disrupting-energy-industry/. [Accessed: 13-Jul-2018].

[65] R. G. Junker, A. G. Azar, R. A. Lopes, K. B. Lindberg, G. Reynders, R. Relan, and H. Madsen, "Characterizing the energy flexibility of buildings and districts," *Appl. Energy*, vol. 225, no. May, pp. 175–182, 2018.

[66] M. Hansen, "Smart grid development and households in experimental projects," pp. 1–125, 2016.

[67] G. E. Videre, "8. januar 2016 | Baggrundspapir til Giv Energien Videre | Udbyg vedvarende energi i takt med efterspørgslen | Side 0 |," pp. 0–25, 2016.

[68] EnergiFyn, "Components of the Electricity Price in Denmark (Sådan er elprisen sammensat)," *EnergiFyn*, 2018. [Online]. Available: https://www.energifyn.dk/privat/elhandel/bag-om-elprisen. [Accessed: 18-Jul-2018].

[69] | T. H., "Danskerne betaler igen rekordlav elpris med et meget stort men," *Mandag Morgen*, 2017. [Online]. Available: https://www.mm.dk/videnbank/artikel/elpriser2016. [Accessed: 18-Jul-2018].

[70] J. M. Rasmussen, "Faste afgifter bremser energibesparelser," *Bolius*, 2014. [Online]. Available: https://www.bolius.dk/faste-afgifter-bremser-energibesparelser-21064/. [Accessed: 14-Nov-2018].

[71] T. Warren Liao, "Clustering of time series data - A survey," *Pattern Recognit.*, vol. 38, no. 11, pp. 1857–1874, 2005.

Part II – Papers 1-4

Paper 1 - Structured Literature Review of Electricity Consumption Classification Using Smart-Meter Data

energies

MDPI

Review

Structured Literature Review of Electricity Consumption Classification Using Smart Meter Data

Alexander Martin Tureczek * and Per Sieverts Nielsen

Systems Analysis, Management Engineering, Technical University of Denmark, 2800 Kongens Lyngby, Denmark; pernn@dtu.dk

* Correspondence: atur@dtu.dk; Tel.: +45-2346-0989

Academic Editor: Mark Deinert

Received: 22 February 2017; Accepted: 19 April 2017; Published: 25 April 2017

Abstract: Smart meters for measuring electricity consumption are fast becoming prevalent in households. The meters measure consumption on a very fine scale, usually on a 15 min basis, and the data give unprecedented granularity of consumption patterns at household level. A multitude of papers have emerged utilizing smart meter data for deepening our knowledge of consumption patterns. This paper applies a modification of Okoli's method for conducting structured literature reviews to generate an overview of research in electricity customer classification using smart meter data. The process assessed 2099 papers before identifying 34 significant papers, and highlights three key points: prominent methods, datasets and application. Three important findings are outlined. First, only a few papers contemplate future applications of the classification, rendering papers relevant only in a classification setting. Second; the encountered classification methods do not consider correlation or time series analysis when classifying. The identified papers fail to thoroughly analyze the statistical properties of the data, investigations that could potentially improve classification performance. Third, the description of the data utilized is of varying quality, with only 50% acknowledging missing values impact on the final sample size. A data description score for assessing the quality in data description has been developed and applied to all papers reviewed.

Keywords: smart meter; data analysis; classification; review; electricity consumption classification; consumption classification

1. Introduction

Recent developments in digital intelligent smart meters have made it possible to monitor energy consumption in details never before seen. The intelligent meters are part of a digitized society, which has been introduced over the last two decades, wherein home appliances, home automation and the smart meters make it possible to monitor energy consumption down to the second. Historically we have measured energy consumption at the household level with analog meters installed at every consumer, and biannually the consumer has reported the meter reading to the utility company for billing purposes. Conversely, intelligent meters are directly connected to the utility company and are able to measure consumption autonomously down to seconds, made possible by the technological development and also pushed by legislation. The intelligent meters enable fast and accurate billing, and also offer a unique and unprecedented opportunity to log and analyze electricity consumption at the consumer level.

Across the European Union member states have initiated installation of the intelligent meters. The European Commission sees the installation of smart meters as a way to improve the overall efficiency of the energy system, and the target is to reach 80% roll out by 2020 in the EU, with an expected reduction in CO_2 emissions by 9% [1]. Denmark has passed legislation that requires all electricity consumers in Denmark, more than 3 million households and industries, to have intelligent

meters installed by the end of 2020. The meters must record consumption at a frequency of no less than 15 min; a level of monitoring that yields at least 35,040 measurements per meter per year.

The high frequency electricity consumption data contain detailed information about consumption patterns, and this has initiated discussions among energy system stakeholders about utilizing the data for purposes other than billing. It has sprawled diverse research projects; such as research on data security and anonymization, non-intrusive load monitoring, load forecasting and consumer classification.

With this in mind, the purpose of this paper is twofold. First; to apply a modified version of Okoli's structured literature review process for conducting an extensive and structured review of smart meter consumption classification, and second; to evaluate the current state of the art in electricity consumption classification using smart meter data and how these findings have been utilized. The review specifically identifies datasets, applied classification methods, results and potential gaps in the research into consumption classification using smart meter data. Papers assessed in this review apply smart meter data in the context of classifying electricity consumption.

Relevant papers have been identified using Thomson-Reuters Web-of-Science search engine, which was selected because of fast search options across multiple scientific journals with multiple search phrases. Thirty phrases were applied in the search and reported in this article. Although this review will not constitute an exhaustive list of search phrases or relevant papers the structured approach encompasses and identifies the most important contributions to the field of electricity customer classification using smart meter data.

This paper will adhere to Fink's [2] definition of systematic literature review: "Such a review must be systematic in following a methodological approach, explicit in explaining the procedures by which it was conducted, comprehensive in its scope of including all relevant material, and hence reproducible by others who would follow the same approach in reviewing the topic" [3]. Even though Web-of-Science indexes many of the leading journals, there will always be papers that are not included in the database or simply do not comply with the selected search phrases. Despite this the approach will present a strong structure and a strict methodology, and encompass the key features and work in this research field, while maintaining reproducibility.

This review will identify the state of the art for electricity customer classification using smart meter data. It will identify methods and datasets, but it is outside this paper's scope to describe the methods identified. The paper is structured as follows: Section 2 introduces the systematic review processes as suggested by Okoli [3], including a practical case on smart meter data in Section 3; Section 4 synthesizes the findings from the structured review process; and Section 5 discusses the findings and the future perspective of the research.

2. Systematic Literature Review Methodology—An Empirical Study

Okoli [3] stresses the importance and difference of systematic reviews versus conventional literature reviews: "rather than providing a base for the researcher's own endeavours it creates a solid starting point for all other members of the academic community interested in a particular topic" [3], and that they are "studies that can stand on their own, in themselves a complete research pursuit" [3], with the "distinguishing feature of a stand-alone review is its scope and rigour" [3]. The point is that the systematic literature review has to be completed with a rigor and systematism that enable others to reproduce the work using the exact same approach. He emphasizes the importance of this type of review, as it represents a base for the community to summarize the bulk of knowledge on the topic. Okoli presents an eight-step guide to conduct systematic literature reviews in information sciences. This paper slightly modifies the original methodology by combining data extraction and synthesis into one category that better fits quantitative studies where the extraction and synthesis of knowledge are closer linked, compared to qualitative studies. The seven steps of the process are outlined below:

Modified Okoli process for systematic literature review:

1: Purpose of the literature review. Clearly state the purpose of the review. What is the scope and contribution of the work presented?

2: **Protocol and training.** Ensure consistency, alignment, and reproducibility by formally defining rules and evaluation criteria.

3: **Searching for literature.** Explicitly describe the search for literature search, the "what and where."

4: **Practical screen.** Crude inclusion and exclusion of articles not based on quality appraisal but on "applicability to the research question." The reviewer normally only reads the title and abstract at this stage. "The practical screen is to screen articles for inclusion. If the reviewer thinks that an article matches the superficial qualities of the practical screen it should be included" [3]; if in doubt the article should be included.

5: **Quality appraisal.** Screen for exclusion, and explicitly define the criteria for judging articles. All articles need to be read and scored for their quality, depending on the research methodologies employed by the articles [3].

6: **Data extraction and synthesis of studies.** Systematically extract the applicable information of the identified articles and combine the facts.

7: **Writing the review.**

In the following sections the method will be applied in an empirical study of "electricity consumption classification using smart meter data." Section 2 of this paper encompasses step 1 and 2, stating the purpose and protocol. Section 3 describes steps 3–5: searching for literature, screening and quality appraisal of the selected papers. Section 4 will address the data extraction and synthesis from step 6, followed by Section 5, where results are discussed.

2.1. Purpose of the Literature Review

The purpose of this paper is to create a systematic literature review of electricity consumption classification using smart meter data. The review will apply a modification of the described systematic literature review process as the basis for a structured and reproducible review, identifying important contributions to electricity consumption classification research. The review will identify significant datasets and methods for classification, point out common denominators and highlight research gaps. The result is an extensive overview of what has been done in the field of smart meter consumption classification and what the authors see as the next step in applying smart meter data.

This review only includes peer-reviewed papers employing electricity consumption data for classification. Research into identification of specific appliances such as Non-Intrusive Load Monitoring (NILM), data collection systems and protocols, smart meter control and development, data privacy and tariff development are beyond the scope of this paper. Only papers published in English are included in this review to maintain reproducibility, fully acknowledging the quality of non-English research literature. The use of the English language is extensive in science and will encompass the current state of the art in smart meter classification.

2.2. Protocol and Training

Regardless of the number of reviewers working on a review it is advisable to develop a formal protocol with evaluation criteria for inclusion, exclusion and quality appraisal to ensure consistency across the reviewers and papers. For this paper a protocol was developed for evaluating and extracting data.

3. Article Selection

The following section will describe how the relevant literature was selected and screened. Section 3.1 describes the search for literature; Section 3.2 describes the initial crude inclusion and exclusion on title and split on paper topic. This is extended in Section 3.3 through screen of abstract,

while Section 3.4 describes the final selection of papers for this review and the quality appraisal. The entire section is equivalent to steps 3–5 in the modified Okoli process.

3.1. Searching for Literature

There are several ways to search for literature. Popular and feasible strategies are to visit multipurpose search-engines like Google, Bing etc. or visit the academic publishers' online resources and identify journals of interest, but as many journals are cross disciplinary it is not a simple task to identify relevant journals. This review uses the Thomson-Reuters search engine Web-of-Science (WoS), which is a comprehensive search engine for academic literature from books, conferences, symposiums and journal papers. It enables the user to search in topic, title, abstract, author etc. across more than 12,000 scientific journals [4].

The WoS engine was set up to search in title or topic. Thirty search phrases with relevance to smart meter data analytics were employed to identify relevant literature. Twenty-six phrases start with the words "smart meter" plus an amendment, such as "*classification*" resulting in the search phrase "*smart meter classification*". Additionally, "*electricity customer classification*", "*electricity customer segmentation*", "*residential electricity classification*" and "*residential electricity segmentation*" could potentially be relevant and include smart meter data. These four search phrases were also integrated into this review even though they do not contain the "smart meter" prefix, so there were 30 search phrases in total. The complete list of search phrases employed is listed in the Appendix A.

3.2. Screening I: Title

Via WoS the 30 phrases resulted in 3922 papers, of which 2099, or approximately 53.5%, were unique. The title of each paper was read and marked for potential relevance for this study; if in doubt the paper was included. The metadata were compiled into an Excel sheet with their relevance marked along with the date of extraction. Due to the manual workload, the extraction took place over seven days, from 5–12 July 2016. Metadata for all papers included and excluded were kept on record.

As previously mentioned papers were excluded for a wide variety of reasons. Every paper remotely relevant to research on electricity consumption classification using smart meter data is included for further screening. After the initial search and screening of the 2099 unique articles, 552 were deemed related and relevant.

3.3. Screening II: Abstract and Removal of Non-Peer-Reviewed Papers

The second screening evaluates the abstract to give a clearer understanding and deeper assessment of the focus of each paper and establish relationships to smart meter data analytics. Each paper was on the basis of the abstract labeled according to its topic, resulting in 10 different topics shown in Figure 1.

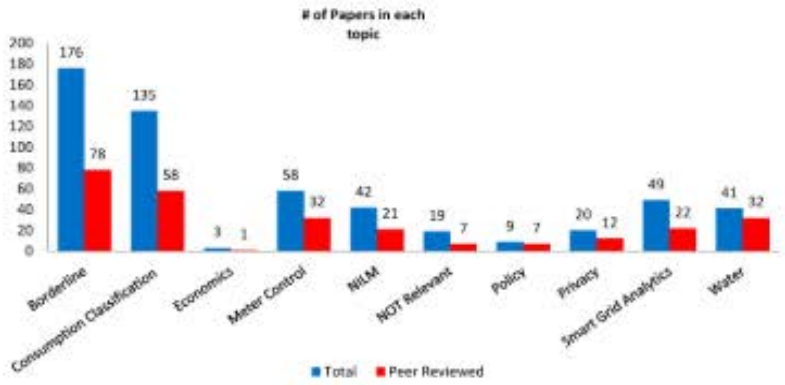


Figure 1. Category distribution after abstract screen. Total includes all types of written material from papers, conferences, workshops etc. Peer-reviewed only include journal papers.

The label **Borderline** (176) papers are potentially relevant for this review; it is not possible from the abstract to conclude if the papers utilize smart meter data or not. **Consumption Classification** (135) through application of smart meter data. **Economic** (3) papers are concerned with grid level business models. **Meter Control** (58) is research regarding smart meter development, control systems and data management. **Non-Intrusive Load Monitoring** or NILM (42) studies how to identify individual appliances and other electric components operated in households through application of smart meter data. **Not Relevant** (19) papers are concerned with health meters, transmission protocols and standards and do not necessarily utilize smart meter consumption data. **Policy** (9) papers address issues about tariff policy and qualitative behavioral studies. **Privacy** (20) papers are focused data security and privacy. **Smart Grid Analytics** (49) are related to the entire distribution and transmission infrastructure. **Water** (41) applies smart meter readings to water consumption. Figure 1 shows the distribution of the distinct categories and the number of peer-reviewed material in each.

Only papers from the research topics **Consumption Classification** and **Borderline** were included in this review. Abstracts from Consumption classification indicate that smart meter data are utilized for classification while borderline can contain papers that apply smart meter data for classification. The papers selected need to be read for quality appraisal to conclude if they are relevant for the review. For quality assurance only papers listed as peer-reviewed journal papers were included in the bulk of relevant papers. Though the exclusion of conference, symposiums and seminar papers may have deprived this review from including the most up to date ideas and concepts, the task of validating non-peer-reviewed articles was not feasible for this study.

3.4. Quality Appraisal

By only including peer-reviewed papers the number of papers was reduced to 58 '*Consumption Classification*' papers and 78 '*Borderline*' papers, adding up to 136. Borderline was revisited by screening all papers for dataset description, resulting in 13 papers applying smart meter data. The 13 papers from Borderline were potentially relevant resulting in a total of 71 papers.

71 papers were read, with special focus on data description, methodology and purpose. Of the 71 papers, 34 focus on clustering consumption; these 34 papers are included in the synthesis of studies. Appendix B includes a qualitative summary table of data extracted from the papers, and Appendix C includes a list of the 34 papers analyzed. A waterfall statistic depicting the screening impact on the final number of papers included in the review can be seen in Table 1.

Table 1. Waterfall statistics showing how many articles were excluded in each step of the screening process.

Waterfall Statistics	Bulk	Reduced
Initial	3922	-
Unique	2099	1823
Screening I: Title	552	1271
Screening II: Abstract	311	241
Removal of non-peer-reviewed	136	175
Screening III: Borderline revisited	71	65
Screening IV: Reading of articles	34	37
Final number of papers synthesized	34	-

4. Data Extraction and Synthesis of Studies

The focus of the 34 selected papers is classification. Many different classification techniques have been tested on smart meter data. Dimensionality reduction has also been applied in order to make large data sets computationally feasible or ease visual inspection. Cluster indices have been applied to evaluate the stability of the resulting clusters. Generally, a large effort has been put into thorough description of methods for classifying consumption and validating the results using smart meter data. Surprisingly the description of the applied data does not adhere the same standard. The following chapter will describe the extracted information of the 34 articles and is divided into

4 sections. Section 4.1 discusses data description and introduces a 13-step data quality score. Section 4.2 is concerned with data classification techniques, while Section 4.3 focuses on dimensionality reduction and feature extraction. Section 4.4 describes the applied validation techniques for ensuring consistency in the clustering. This section complies with step 6 in the modified Okoli process.

Table 2 summarizes descriptive empirical information regarding the origin of data, how long the data have been recorded, at what frequency, the number of meters available and some of the classification methods applied.

4.1. Data Description and Empirical Findings

An important part when working with data analytics is knowledge about the data. This knowledge must be conveyed such that the reader gets an understanding of the data and how it can be utilized for analysis. For smart meter data, such information is sample size, supplier and customer type; residential or industrial. The 34 selected papers in this review demonstrate varying attention to these details when describing the data used in their research, some papers invest great effort while others apply much less care describing the data.

In order to quantify the quality of data description in each paper, the authors have created a data quality score, which is comprised of 13 measurable attributes. An attribute identified in the paper adds to the score, for a maximum of 13, the attributes are uniformly weighted.

The 13 attributes create a baseline of insight into the data used in the paper. The very thorough qualitative description of data seen in [5,6] elevates the level of description from the baseline but it is not honored in this score. The score is intended as a checklist for essential information when describing electricity smart meter data, and there are 5 categories comprising the score: Geographical information, Data information, Time information, Type information, References.

Geographical information (3 points): The **country** where the data was collected is relevant to assess possible (de)similarities in consumers and energy systems. **Region** is relevant for the understanding of the consumers, is the region. Is the region scarcely populated, having fluctuating climate or other identifying features? **Supplier** indicates who supplied the data for the study, and describes how representative it is of the population.

Data information (4 points): the initial **size** of the dataset is very relevant to reader and the generalizability of the results. Any real-life data set needs preprocessing before it is applicable for analysis, was certain consumers removed from the data, or were there other exclusion criteria? There should be a **clear description of the reduction** this preprocessing had on the sample size. After preprocessing is there listed an unambiguous **final sample size**? The data is generated from meters which are prone to random errors or **missing values** in the recordings. Have the authors acknowledged data imperfections and included a description of how missing or erroneous recordings were resolved.

Time information (4 points): The recording interval has a significant impact on the analytical challenges the data can help explain therefore the **recording frequency** must be stated including **commence** and **end** of the recordings**.** The **length** of uninterrupted recordings gives an indication of generalizability and the possibility of doing classification on daily, monthly, quarterly or yearly data.

Type information (1 point): The **type** of consumers, residential, industrial or both, the data set includes. These clients can exhibit vastly different consumption patterns.

Referencing other data sets (1 point): A Paper can reference data description in other papers of the same data. This attribute has been included to enable articles without data description to get acknowledgement through other papers describing the exact same data set. This is also relevant if the authors have described data in a previous paper. If the attribute information exists in the referenced papers these are counted in the score.

The developed data description score has been applied to all 34 papers in this review. For illustration purposes Table 3 shows the scoring of two papers. Both score 12 but they don't include the same information; paper [7] has no data referral while paper [8] has no mentioning of the supplier of the data.

Table 2. Data and method overview table showing the quantity of different datasets encountered in each country and the main attributes from the data description score, Country, Type, length, frequency, final sample size, and the most prevalent classification encountered in this review. Denmark has no method listed as the classification used is a combination of regression and survey data.

# of Different Datasets	Country	Type	Length	Recording Frequency	Final Size (m)	K-Mean	Hierarchical	Follow the Leader	Mixture Models	Fourier Transform
1	Brazil	Mix	1 month	15 min	2000	x	-	-	-	-
2	Canada	Mix	12 months	60 min	62,923	-	-	-	x	
1	Denmark	Mix	36 months	60 min	4500	-	-	-	-	-
1	Finland	Mix	12 months	60 min	3989	x	x	-	-	-
1	France	Residential	UNKNOWN	10 min	100	x	-	-	-	-
1	Germany	Residential	UNKNOWN	15 min	215	x	-	-	-	-
1	Greece	Industrial	10 months	15 min	292	x	x	-	-	-
1	Ireland	Residential	>24 months	30 min	3941, 3440, 3622, 3487	x	-	x	x	-
1	Japan	Residential	18 months	1 min	1072	x	x	-	-	-
3	Korea	Mix	27 months	15 min, 60 min	1735, 1205, 135	x	x	-	-	-
2	Portugal	Residential	48 months	15 min	265, 1022	x	x	-	-	-
1	Romania	Industrial	1 day	15 min	234	x	x	x	-	-
2	Spain	Mix	24 months	60 min	711, 230	x	-	-	-	x
2	UK	Residential	18 months	30 min	5000	x	-	-	-	-
1	UK, Bulgarian	Residential	12 months	7 s	197	x	x	-	x	-
5	USA	Mix	>8 months	1 min, 10 min, 60 min	952, 123,150, 952, 103, 2000	x	x	-	-	-

Table 3. Example of the Data Description Score result on 2 articles, paper 7 has no referral and scores 12, while paper 8 has referral but no Provider information and equally scores 12.

Reference Article #	[7]	[8]
Country	x	x
Region	x	x
Supplier	x	-
Initial Size	x	x
Clear Reduction	x	x
Missing Values	x	x
Final Size	x	x
Recording Frequency	x	x
Start	x	x
End	x	x
Length	x	x
Type	x	x
Referral	-	x

Table 4 shows the penetration of different attributes in the papers, and no attribute is accommodated by all. The most prevalent are identifiable in 33 (97%) papers. There is a consensus among more than 30 (>90%) papers that country, initial size, clear description of reduction, final size, recording frequency and consumer type is relevant information to state in a paper when describing the data 30 papers (88%) found that the length of recording is essential information to state when describing the data, while 27 papers (79%) include information about the region and supplier of the data. Only 23 respectively 22 found it important to include information about commencing and end time for the recording.

Table 4. Data Description Score comprising of 5 main categories with a total of 13 different attributes. Prevalence describes the number occurrences of the attribute in the 34 papers. Percent shows the prevalence percent of the 34 papers.

Category	Attribute	Prevalence	Prevalence%
Geographical information	Country	33	97%
	Region	27	79%
	Origin	27	79%
Data Information	Initial Size	33	97%
	Clear Reduction	32	94%
	Missing Values	17	50%
	Final Size	31	91%
Time Information	Recording Frequency	33	97%
	Start	23	68%
	End	22	65%
	Length	30	88%
Type Information	Type	32	94%
Referencing Data	Referral	13	38%

It is surprising that missing values is only mentioned in 50% of the papers. Missing data is prevalent in any real-life datasets, and how they are resolved is important to describe to account for any bias. The description can be very brief, and paper [9] shows how a short yet detailed description of data preprocessing, with encountered issues and main strategies for alleviating them, can be integrated into a paper.

Table 5 shows the distribution of scores by the 34 papers. It is seen that 3 papers have a mentioning of all 13 attributes, while 23 papers include 10–12 attributes, resulting in 26 papers scoring 10 or more.

Table 5. Distribution of papers for different scores.

Score	Quantity
7	4
8	1
9	3
10	8
11	8
12	7
13	3
Grand Total	34

4.2. Classification

With more than 10 different classification methods applied in 34 papers, the most prevalent methods observed was K-means clustering. K-means and related methods like K-medoid and Fuzzy K-means are used in 22 (65%) articles, often for performance comparison to more advanced

techniques [10–12]. The popularity of K-means clustering can be attributed to its simplicity and generally satisfactory performance. K-means is also implemented in many software solutions, both proprietary and open source, making it an easy choice for fast clustering. The greedy design approach of the K-means algorithm can create suboptimum solutions by unfortunate initial starting conditions and converge in local optima; a problem that can be alleviated by rerunning the algorithm several times [13,14].

Agglomerative hierarchical clustering is used in 10 (29%) of the papers. Hierarchical clustering offers intuitive graphical display and interpretation of the class evolution for different thresholds in one figure. This method requires considerations about distance measures given by the link function. Popular link functions are the Euclidian, Wald and average linking.

More advanced models like Follow-the-leader and Mixture models are observed respectively in 5 (15%) and 3 (9%) papers for instance in [15,16]. The clustering is frequently applied directly on the raw data without investigating inherent features in data that could aid in classification, features like autocorrelation, seasonality, variance and average. Many features can automatically be extracted from data, dimensionality reduction techniques such as principal components or self-organizing maps can help identify hidden features.

Smart meter data can be regarded as signals; as such it could be advantageous to apply techniques that leverage time series information like periodicity or autocorrelation. In [12] Fast Fourier transform (FFT), a frequency domain analysis technique for signals, is applied, but several other techniques exist for analyzing time series. Wavelet transform, a signal processing method, is good for feature extraction and dimensionality reduction, and could be an interesting addition to the analysis of smart meter data but was not encountered in any of the papers.

The most frequent approach to customer classification encountered is unsupervised learning. In [17] artificial neural networks are applied for supervised learning by mapping data to clusters in an input to response ($y = a \cdot x$) manner, but this cannot be done without prior knowledge of the clusters. To identify the clusters K-means is applied for unsupervised clustering before creating a neural network. Regression techniques like Hidden Markov models [18], linear regression [7] or logit [19], [20] are utilized for supervised classification of consumption, but are all using unsupervised clustering or survey data for initial starting conditions. Table 6 show a summary of the most frequently encountered methods with their most notable properties, while Table 7 gives an overview of different link functions applied in for instance Hierarchical clustering.

Table 6. Overview of the most prevalent classification methods encountered in the 34 papers.

Method	Advantages	Caveats
K-means	Fast, well documented.	Risk of Local optimum. Difficult to find optimum cluster number and interpret clusters.
Hierarchical	Visual interpretation, fast.	Careful selection of link function is required.
Follow-the-leader	No initial number of clusters to fit.	Needed Distance threshold is chosen by trial-and-error
Mixture Models	Advanced modelling of systems.	Complex setup compared to K-means
Neural Network	Supervised, taking into account prior knowledge.	Risk of overfitting, needs prior knowledge

Table 7. Overview of the most prevalent distance measure encountered in the 34 papers, and their behavior.

Link Function	Distance between Clusters	Cluster Behavior
Single (Euclidian)	Closest	Long non-convex cluster shapes
Average	Average	-
Complete	Largest	Convex clusters, sensitive to outliers
Centroid	Cluster center	Robust vs outliers
Ward	smallest variance	Equally sized clusters

4.3. Dimensionality Reduction and Feature Extraction

Unsupervised classification techniques like K-means do not consider the inherent information stored in time series. There is no connection to autocorrelation or other features in the data. The algorithm regards every time-step as a feature or a dimension with no correlation to neighboring readings. Not necessarily a problem but with long recording windows—weeks or months—and few meters the curse of dimensionality could impact the applicability of the results. The curse revolves around increase in the required amount of data when increasing the number of dimensions; this is an exponential growth pattern and can render the data insufficient for the analysis.

Real life data—regardless of dimensions—often have some natural clustering or dependencies which can be highlighted and exploited by dimensionality reduction [21]. A popular algorithm for reducing dimensions in smart meter data sets is Self-Organizing Maps (SOM). SOM projects data into lower dimensions and can be useful for visualization, "SOM is an algorithm characterized by robustness and computational efficiency" [22]. While [23] notes that SOM is useful for handling noisy data and outliers due to the dimensionality reduction, which in turn results in better performance from K-means and other clusters algorithms compared to direct application of the algorithms on raw data.

SOM also delivers unsupervised classification which "can be viewed as a constrained version of K-means clustering in which the prototypes are encouraged to lie in a one or two dimensional manifold of the feature space" [24]. The two-dimensional manifold gives SOM desirable properties for visual inspection of the data.

From the papers, it is inconclusive whether to apply dimensionality reduction on a smart meter data set; "In general, the counterpart of the benefits of data size reduction is lower classification effectiveness, in terms of higher clustering validity indicators. On the basis of the results, the validity of the data size reduction methods can be generally indicated as acceptable" [10], rendering the application of dimensionality reduction at the discretion of the researchers on case by case basis. It is a trade-off between classification effectiveness and validity of clusters.

4.4. Cluster Validity Check

Estimating the optimum number of clusters is not a trivial task. Without prior knowledge of the underlying clusters there is no unambiguous way to identify the true underlying clusters. In an effort to quantify the uncertainty of the clusters [10], applies 4 different indices for cluster evaluation, of which the Davies-Bouldin (DBI), Mean Index Adequacy (MIA) and the cluster-dispersion index (CDI) are frequently applied in other papers for cluster selection.

Where regression is a minimization problem, minimizing sums of squares minimizing variance in clusters would yield the same number of clusters as meters which is not desirable. A wide array of indices for validation of cluster stability has been developed to aid the cluster selection process. There is no shortage of validity indices; the 34 papers in this review employ 18 different indices [25], notes: "Although these indexes [DBI, CDI, MIA] are widely accepted in clustering, they are not proficient in specific applications such as electricity load profile clustering. They do not consider domain knowledge and only focus on the internal structure of nodes. For these reasons, we focus on an external validity index such as entropy, which compares the clustering answer with pre-assigned, ground-truth clusters". There still does not exist a single adequate index for validation of clusters, as with model diagnostics in regression the combination of indices help give an overview of the performance. Table 8 lists the most prevalent indices in this review and lists their properties.

Table 8. Popular cluster validation indices and how to interpret them.

Index	Mathematics	Interpretation
DBI (Davies-Bouldin Indictor)	$\frac{1}{K}\sum_{i=1}^{K}\max_{j\neq i}\frac{diam(C_i)+diam(C_j)}{d(C_i,C_j)}$	$diam(C_k)$ is the average diameter of a cluster. And $d\left(C_i,C_j\right)$ is the distance between centroids. K is the number of clusters. DBI relates the mean distance of each class with the distance to the closest class [26]. Smaller values of DBI implies that K-means clustering algorithm separates the data set properly [23]
CDI (Cluster Dispersion Indicator)	$\frac{1}{d(C)}\sqrt{\frac{1}{K}\sum_{k=1}^{K}d^2(C_k)}$	CDI prefers Long inter-cluster distance and short intra-cluster distance [25]. Small values indicate good clustering. $d^2(C_k)$ is the squared average distance within cluster k. High. While $d(C)$ is max cluster distance in data.
Dunn	$\frac{\min d(C_i,C_j)}{\max diam(C_m)}$ where $d\left(C_i,C_j\right)=\min_{x\in C_i,y\in C_j}\|x-y\|$ and $diam(C_k)=\min_{x,y\in C_k}\|x-y\|$	The ratio between "minimum distance between clusters" and "maximum distance within clusters". When minimum dissimilarity between clusters get large and max cluster diameter gets small the Dunn value gets large and indicates good separation. C_i is cluster i, d is distance and m is total number of clusters.
Silhouette	$\frac{c'(x)-c(x)}{\max\{c(x),c'(x)\}}$ $c'(x)=\min_{y\in C'}d(x,y)$	$c(x)$ is the average distance between vector x and all other vectors of the cluster c to which x belongs. $c'(x)$ is the minimum distance between vector x and all other vectors in cluster $\forall\, C'\neq C$ [23]. SI is between [–1, 1] higher is better. Negative is miss-clustering.
Entropy	$-\sum_{i=1}^{c}p\left(\frac{i}{t}\right)\cdot log_2 p\left(\frac{i}{t}\right)$	$p\left(\frac{i}{t}\right)$ denotes the proportion of correct classified vector i in cluster t. Entropy is a supervised index as the true classes needs to be known. Entropy is used as a measure of misclassification in each cluster. Entropy is small when the clustering result is similar to the expected result [25]. c is total clusters.
MIA	$\sqrt{\frac{1}{K}\sum_{k=1}^{K}d^2(C_k)}$	Average distance within class to class centroid, summarized across all classes. k is number of clusters; $d^2(C_k)$ is the squared average distance within cluster k. High MIA indicates large distances within the classes. e.g., large dispersion.

5. Findings and Discussion

The proposed method for conducting a structured literature review applied in this paper has supplied a structure and simple step-by-step guidelines for ensuring consistency, objective evaluation and selection of papers. For a detailed summary of the qualitative findings from the 34 analyzed papers, see Appendix B.

In unsupervised segmentation, no prior information exists about the true underlying classes. There is no unambiguous minimization problem that can identify the true clusters. To alleviate the difficulties in selecting the number of clusters the literature has developed a wide variety of cluster performance estimators. The performance estimates help researchers determine the optimum number of clusters. Performance estimators evaluate different information [25], as with unsupervised classification, the performance estimates should be perceived as a tool to validate and not prove the correctness of the clusters.

The evaluated papers roughly follow the modelling structure outlined in Figure 2. Blue indicate the elements all papers undergo; Describing the **data** from meters and applicable external data. **Method** selection, dim reduction and classification algorithms. **Clustering** of the meter data and **validation** of the clusters to select optimum classification [10].

Successful classification leads to interest in cluster composition. This is often done with the aid of external data such as [6,27] combine smart meter data with survey data to attain deeper knowledge of the identifying features of the individual clusters.

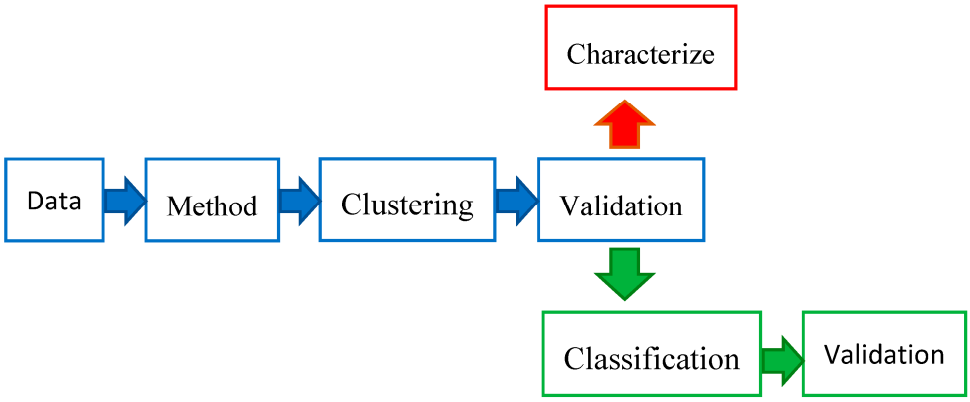


Figure 2. Depiction of standard modelling structure. (Blue) Indicates the elements all papers go through (data, method selection, clustering and validation); (Red) Some papers charachterize the identified clusters, usually applying external data; (Green) Applying the identified clusters to classifify new consumption series to evaluate the applicability of the clusters.

Another application of the clusters is applying the clusters to new data in order to test the classification abilities of the clusters identified (**green**) on unknown data. Clustering on new data requires validation of the resulting classification a process similar to validating the initial clulsters.

Few papers charachterize the entire process of clustering, characterization and classification of new data [20], for a more complete overview of consumption behavior. While papers focusing only on clustering and validation (**blue**) result in more detailed comparissons of cluster methods.

Some papers are focused on evaluating unsupervised classification and cluster validation techniques [10,13,16], while others are looking at the implications of the clusters on our understanding of consumption.

Few papers also characterize the clusters identified (**red**). A Portuguese study with 265 meters enriched with survey data showed that it is possible to segment into distinct clusters and make meaningful socio-demographic deductions about the different clusters [6]. Through the clusters the study recognized 3 types of consumption: "fuel poverty" as households not keeping their home adequately warm, "Standard comfort" households and "fat energy" households which could be more rational in their consumption pattern. A similar study in Japan shows the ability to identify different consumption patterns and quantify the excess energy used for different life styles only by analyzing smart meter data, and discuss how these results can be used to influence the residential consumption by targeted and personalized information [12]. The Japanese study shows how differences in daily routines influence the consumption, while the Portuguese study identifies distinct levels of consumption within the same daily routines.

The high frequency time series created by smart meters, give invaluable insights into electricity consumption. For instance, the meters make it possible to investigate how well the UK Elexon profiles fit modern data segmentation techniques. The Elexon profiles divide all electricity customers in the UK into seven distinct clusters, two for residential and 5 for industry. "The usefulness of these Elexon "profiles" for domestic customers is unsatisfactory. It has been reported that the use of the profiles has made about 9×10^{12} W·h electricity losses yearly in the UK" [14]. Smart meter analytics will become increasingly important in the development of the electricity grid through consumption insights. Consumption profiles can aid in the construction of dynamic tariffs for more fair pricing [28] and smarter utilization of the existing grid by creating economic incentives for consumption flexibility. Survey and socio-economic data can further improve the understanding of consumers and optimize the electricity grid. It's a good business case reducing costs while simultaneously reducing the carbon footprint from electricity production.

Some papers identify yearly seasonality, and are also able to identify distinct time periods during the week or day [29]. The inclusion of weather information is rare and when included it is for improving the model with temperature compensation [18,22]. Weather compensation is generally applied when the meter readings are collected in different regions or with high seasonal variations with varying consumption across the year.

Smart meter data evolves over time and can to some extend be expected to contain autocorrelation. Few of the papers apply time series techniques to leverage this. K-means evaluates time steps independently, and does not account for any correlation structure between the time steps. It is a fast and capable method which is implemented in all major statistical packages and simple to implement. It has drawbacks, such as getting trapped in local optima and not leveraging the correlation in the data. To account for autocorrelation or multicollinearity in the data, methods for dimensionality reduction are applied. Principal Component Analysis or Self-Organizing Maps removes correlation structures and maps the data to a reduced feature space, but it comes at a cost of interpretability of the final results. Paper [12] applies time series techniques through Fourier transformation of the meter data which results in a frequency spectrum for each meter, then applying K-means on the largest peak in the spectrum. Fast Fourier Transform stems from signal analysis; it converts data from time domain to frequency domain and lists the observed frequencies, which then can be considered as features. In the Fourier transform one would have to decide between time or frequency domain as there is no interpretable link between them. You would know the frequencies but not when they occur. Interestingly no paper looked at classical time series with ARIMA models and how to classify them. Finally, it would have been interesting to see Wavelets applied as they combine time and frequency information in contrast to Fourier transform. Furthermore, Wavelets are capable of reducing dimensionality and extracting features, which can be used as input in K-means classification.

The analysis of papers included in this review show that they vary greatly in the effort put into describing the data applied in their research. Most include information regarding country, supplier and recording frequency, prevalent in 33 out of 34 papers. Surprisingly only 50% of the papers report any information on encountering and handling missing values. The meters producing the time series can be subject to random issues with transmitting data, resulting in missing meter readings that need to be rectified. Some missing values can be imputed while others are more severe and need the entire meter series to be discarded from the study. If values were imputed, a clear description of the processes is needed to evaluate the implications. Discarding has an immediate effect on the final sample size. Both processes have implications on the data and the resulting analysis and deductions, and lack of description impacts reproducibility of the study. It is expected that any data set will contain imperfect data, and it is surprising that so many papers neglect to describe or acknowledge this phenomenon. Section 4.1 identifies papers with short and concise description of missing values and remedies which fit into scientific papers.

Over the past decades new household products that run on electricity have been introduced, elevating the individual average consumption of the population. Even though these new appliances improve in efficiency due to technological advances and the electricity consumption in the industrialized world is stabilizing. This stability can be offset by introduction of new technology like computers, or electrification of the transportation sector. Another important component relating to energy consumption is age compositions. In the 34 paper analyzed in this review there is no focus on consumer transition between classes. Classes are regarded as static, derived from data without considering the human behavior they depict can change over time. Suburban areas progress from families with toddlers to teens to elderly until the process repeats every 20–30 year with changing consumption patterns, and this time dependency and the implications of it needs to be investigated. Is it reasonable to assume the clusters identified today are valid in a different setting? Transitions between classes are relevant when planning for future power supply and the insights from smart meter analytics can help identify changes in consumption and transitions, which are valuable for the continued maintenance of the segmentation.

It is not only in households smart meters can have an effect. When electricity replaces fossil fuels in the transportation sector, the demand for electricity will increase substantially as will the variance in the demand. More demand can result in higher peak power through the existing cables, which are designed to cope with a smaller maximum load than the future potentially could demand. It is an expensive and possibly infeasible solution to upgrade the cables to comply with 3–4 times higher peak demand, which electrification of the transportation sector could require. This brings focus to smarter use of the existing grid and the importance of understanding consumption behavior. Peak shaving by moving consumption periods could help alleviate problems with increased demand. This is where smart meters could supply insight and help make the grid and tariffs even smarter.

Denmark projects to reach 84% renewable electricity in 2020 [30], Sweden has aimed high at becoming the first zero-emission welfare country [31]. Zero environmental impact electricity could potentially change consumption patterns to more and different consumption. Will Europeans continue to be energy conscious when there is an abundance of renewable electricity in the grid with no carbon footprint?

6. Conclusions

The proposed method for structured literature review outlined in [3] and demonstrated in this paper has supplied a structure and simple step-by-step guidelines for ensuring consistency, securing objective evaluation and selection of papers. The review applied 30 search phrases with relevance to smart meters, initially encompassing 2099 unique written pieces. Which after extensive screening of title and abstract and the inclusion of peer-reviewed papers, was reduced to 71 papers containing potential studies regarding electricity consumption classification using smart meter data. These 71 papers were thoroughly screened for purpose, data, method and results until a final list of 34 relevant papers concerning electricity consumption classification using smart meter data.

The 34 papers evaluated in this review have shown that electricity consumers are not one homogenous group but can be segmented—using only consumption data—into smaller more homogenous clusters. The clusters are vastly different from the previously used profiles that are built on socio-economic clustering.

Unsupervised learning techniques as the K-means family and hierarchical clustering are widely applied on smart meter data in these papers, either directly for classification or as performance benchmark for evaluation of more advanced methods as follow-the-leader, hidden Markov models and mixture models [10]. For hierarchical clustering the selection of link function influences the clustering performance, several different distance measures are applied.

It is generally concluded that smart meter data is very applicable for cluster analysis, with overall satisfactory performance for individual methods. K-means and hierarchical clustering are simple and fast techniques. While there is some discrepancy in the performance, but all methods introduced can perform satisfactory and meaningful classification of consumption regardless of households or MW consumers.

Having shown that simple classification algorithms like K-means and Hierarchical clustering works on different smart meter data sets we find it appropriate to move focus from simple classification of smart meter consumption data to how these findings provide value in a societal setting. This could be in tariff development or in consumption flexibility analysis. Keeping focus on the statistical classification a more thorough investigation of the statistical properties of the data is a much-needed addition to the standard classification analysis of the data encountered in the analyzed papers. Deeper investigation of the time series properties such as correlation structure and its impact on the classification may contribute to even better understanding of consumption in general.

Acknowledgments: This work is part of the CITIES (Centre for IT–Intelligent Energy System in Cities) project funded in part of the Danish Innovation found. Grant DSF 1305-00027B (Det Strategiske Forskningsråd). The authors wish to thank Pia Thomsen, Rikke Brinkø for helpful inputs, comments and proof reading.

Author Contributions: Alexander Martin Tureczek and Per Sieverts Nielsen conceived and designed the study; Alexander Martin Tureczek performed the literature search and the analysis; Alexander Martin Tureczek and Per Sieverts Nielsen wrote paper.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. List of Search Phrases

Table A1. List of search phrases applied.

Search Phrase	Non-Unique Hits
Electricity customer classification	123
Electricity customer segmentation	34
Residential electricity classification	59
Residential electricity segmentation	16
Smart meter analysis	767
Smart meter analytics	61
Smart meter big data	65
Smart meter classification	187
Smart meter clustering	125
Smart meter consumption classification	47
Smart meter consumption data	704
Smart meter consumption profiling	112
Smart meter consumption segmentation	22
Smart meter customer classification	25
Smart meter customer segmentation	15
Smart meter data analysis	443
Smart meter data mining	46
Smart meter feature construction	12
Smart meter feature extraction	21
Smart meter learning	117
Smart meter load monitoring	262
Smart meter load profiling	147
Smart meter machine learning	47
Smart meter profiling	280
Smart meter segmentation	27
Smart meter statistical learning	6
Smart meter statistics	52
Smart meter supervised learning	7
Smart meter time series	86
Smart meter unsupervised learning	7
Sum	3922

Appendix B. Quantitative Summary Table

Papers reference number in column "Paper(s)" are linked to Appendix C.

Table A2. Data synthesis summary.

Category	Split	Paper(s)
	Classification	
	K-means	[1–23]
	Hierarchical	[1–3,10,15,17,19,20,22,24]
	Fuzzy K-means	[1,15,17,19]
	Follow the leader	[1,9,19,25,26]
	K-medoid	[9,14]
	Mixture models	[10,27,28]
	Fast Fourier Transform	[15]
	Others	[15–21,29]

Table A2. *Cont.*

Category	Split	Paper(s)
	Forecasting	
	Regression	[8,13,27,30–33]
	HMM	[16,23]
	Other	[13,34]
	Dimension reduction	
	Principal Components	[16,34]
	Self-organizing-Maps	[2,9,11,19,21]
	Validation	
	DBI	[1,6,9–11,13,15,17,20,26]
	CDI	[1,10,17,18,22,25,26]
	Dunn & Silhouette	[11,13,14,20]
	Entropy	[19,21,22,28]
	MIA	[6,10,17,25,26]
	Other	[1–3,10,12–18,20–22,26,27,30]
	Size	
	0–250	[1,4,6,8,10,16,18,19,22,25,26,34]
	250–500	[17,24]
	500–1000	[5,23,27,30]
	1000–2500	[7,11,12,15,20,21]
	2500–5000	[2,9,13,14,28,29,31–33]
	Other	[3]
	Region	
	Europe	[1,2,4–6,9,10,13,14,16–20,24–26,28,29,31–33]
	North America	[3,8,12,23,27,30,34]
	Asia	[11,15,21,22]
	Other	[7]
	Period	
	1 day	[1,18,19,26]
	1 month	[7,21,22]
	2–6 months	[9,30]
	7–12 months	[17,23]
	1 year	[2,6,8,10,20,27,28,31,34]
	1–2 years	[3,5,13–15,29,32]
	2+ years	[11,24,33]
	Missing	[4,12,16,25]
	Recording frequency	
	<1 min	[10]
	1 min	[8,15]
	10 min	[16,23,30]
	15 min	[1,4,7,17–20,22,24–26]
	30 min	[9,13,14,28,29,31,32]
	60 min	[2,3,5,6,11,12,21,33,34]
	24 h	[27]
	Type of customer	
	Industrial	[1,12,17–20,22,25,26]
	Residential	[3–5,8–11,13–16,23,24,28–32,34]
	Mix	[2,6,7,21,27,33]

Appendix C. List of Articles

1. Chicco, G.; Napoli, R.; Piglione, F. Comparisons among clustering techniques for electricity customer classification. *IEEE Trans. Power Syst.* **2006**, *21*, 1–7.
2. Räsänen, T.; Voukantsis, D.; Niska, H.; Karatzas, K.; Kolehmainen, M. Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data. *Appl. Energy* **2010**, *87*, 3538–3545.

3. Kwac, J.; Flora, J.; Rajagopal, R. Household energy consumption segmentation using hourly data. *IEEE Trans. Smart Grid* **2014**, *5*, 420–430.
4. Flath, C.; Nicolay, D.; Conte, T.; Van Dinther, C.; Filipova-Neumann, L. Cluster analysis of smart metering data: An implementation in practice. *Bus. Inf. Syst. Eng.* **2012**, *4*, 31–39.
5. Benítez, I.; Quijano, A.; Díez, J.; Delgado, I. Dynamic clustering segmentation applied to load profiles of energy consumption from Spanish customers. *Int. J. Electr. Power Energy Syst.* **2014**, *55*, 437–448.
6. López, J.J.; Aguado, J.A.; Martín, F.; Mu, F.; Rodríguez, A.; Ruiz, J.E. Hopfield—K-Means clustering algorithm: A proposal for the segmentation of electricity customers. *Electr. Power Syst. Res.* **2011**, *81*, 716–724.
7. Macedo, M.N.Q.; Galo, J.J.M.; De Almeida L.A.L.; Lima, A.C.D.C. Demand side management using artificial neural networks in a smart grid environment. *Renew. Sustain. Energy Rev.* **2015**, *41*, 128–133.
8. Rhodes, J.D.; Cole, W.J.; Upshaw, C.R.; Edgar, T.F.; Webber, M.E. Clustering analysis of residential electricity demand profiles. *Appl. Energy* **2014**, *135*, 461–471.
9. Mcloughlin, F.; Duffy, A.; Conlon, M. A clustering approach to domestic electricity load profile characterisation using smart metering data. *Appl. Energy* **2015**, *141*, 190–199.
10. Granell, R.; Axon, C.J.; Wallom, D.C.H. Impacts of raw data temporal resolution using selected clustering methods on residential electricity load profiles. *IEEE Trans. Power Syst.* **2015**, *30*, 3217–3224.
11. Park, S.; Ryu, S.; Choi, Y.; Kim, J.; Kim, H. Data-driven baseline estimation of residential buildings for demand response. *Energies* **2015**, *8*, 10239–10259.
12. Lavin, A.; Klabjan, D. Clustering time-series energy data from smart meters. *Energy Effic.* **2014**, *8*, 681–689.
13. Viegas, J.L.; Vieira, S.M.; Melício, R.; Mendes, V.M.F.; Sousa, J.M.C. Classification of new electricity customers based on surveys and smart metering data Commission for Energy Regulation. *Energy* **2016**, *107*, 804–817.
14. Al-otaibi, R.; Jin, N.; Wilcox, T.; Flach, P. Feature construction and calibration for clustering daily load curves from smart-meter data. *IEEE Trans. Ind. Inform.* **2016**, *12*, 645–654.
15. Ozawa, A.; Furusato, R.; Yoshida, Y. Determining the relationship between a household's lifestyle and its electricity consumption in Japan by analyzing measured electric load profiles. *Energy Build.* **2016**, *119*, 200–210.
16. Basu, K.; Debusschere, V.; Douzal-chouakria, A.; Bacha, S. Time series distance-based methods for non-intrusive load monitoring in residential buildings. *Energy Build.* **2015**, *96*, 109–117.
17. Tsekouras, G.J.; Hatziargyriou, N.D.; Member, S.; Dialynas, E.N.; Member, S. Two-stage pattern recognition of load curves for classification of electricity customers. *IEEE Trans. Power Syst.* **2007**, *22*, 1120–1128.
18. Chicco, G.; Ionel, O.M.; Porumb, R. Electrical load pattern grouping based on centroid model with ant colony clustering. *IEEE Trans. Power Syst.* **2013**, *28*, 1706–1715.
19. Chicco, G.; Sumaili Akilimali, J. Renyi entropy-based classification of daily electrical load patterns. *IET Gener. Transm. Distrib.* **2010**, *4*, 736–745.
20. Ramos, S.; Duarte, J.M.; Duarte, F.J.; Vale, Z. A data-mining-based methodology to support MV electricity customers' characterization. *Energy Build.* **2015**, *91*, 16–25.
21. Piao, M.; Shon, H.; Lee, J.; Ryu, K. Subspace projection method based clustering analysis in load profiling. *IEEE Trans. Power Syst.* **2014**, *29*, 2628–2635.
22. Kang, J.; Lee, J. Electricity customer clustering following experts' principle for demand response applications. *Energies* **2015**, *8*, 12242–12265.
23. Albert, A.; Rajagopal, R. Smart meter driven segmentation: What your consumption says about you. *IEEE Trans. Power Syst.* **2013**, *28*, 4019–4030.

24. Gouveia, J.P.; Seixas, J. Unraveling electricity consumption profiles in households through clusters: Combining smart meters and door-to-door surveys. *Energy Build.* **2016**, *116*, 666–676.
25. Chicco, G.; Napoli, R.; Piglione, F.; Postolache, P.; Scutariu, M.; Toader, C. Load pattern-based classification of electricity customers. *IEEE Trans. Power Syst.* **2004**, *19*, 1232–1239.
26. Carpaneto, E.; Chicco, G.; Napoli, R.; Scutariu, M. Electricity customer classification using frequency—Domain load pattern data. *Int. J. Electr. Power Energy Syst.* **2006**, *28*, 13–20.
27. Coke, G.; Tsao, M. Random effects mixture models for clustering electrical load series. *J. Time Ser. Anal.* **2010**, *31*, 451–464.
28. Haben, S.; Singleton, C.; Grindrod, P. Analysis and clustering of residential customers energy behavioral demand using smart meter data. *IEEE Trans. Smart Grid* **2015**, *7*, 136–144.
29. Tong, X.; Li, R.; Li, F.; Kang, C. Cross-domain feature selection and coding for household energy behavior. *Energy* **2016**, *107*, 9–16.
30. Kavousian, A.; Rajagopal, R.; Fischer, M. Determinants of residential electricity consumption: Using smart meter data to examine the effect of climate , building characteristics, appliance stock and occupants' behavior. *Energy* **2013**, *55*, 184–194.
31. Mcloughlin, F.; Duffy, A.; Conlon, M. Characterising domestic electricity consumption patterns by dwelling and occupant socio-economic variables: An Irish case study. *Energy Build.* **2012**, *48*, 240–248.
32. Jin, N.; Flach, P.; Wilcox, T.; Sellman, R.; Thumim, J.; Knobbe, A. Subgroup discovery in smart electricity meter data. *IEEE Trans. Ind. Infrom.* **2014**, *10*, 1327–1336.
33. Andersen, F.M.; Larsen H, V.; Boomsma, T.K. Long-term forecasting of hourly electricity load: Identification of consumption profiles and segmentation of customers. *Energy Convers. Manag.* **2013**, *68*, 244–252.
34. Ndiaye, D.; Gabriel, K. Principal component analysis of the electricity consumption in residential dwellings. *Energy Build.* **2011**, *43*, 446–453.

References

1. EU Commission. Smart Grids and Meters. Available online: https://ec.europa.eu/energy/en/topics/markets-and-consumers/smart-grids-and-meters (accessed on 20 December 2016).
2. Fink, A. *Conducting Research Literature Reviews: From the Internet to Paper*, 2nd ed.; Sage Publications: Thousand Oaks, CA, USA, 2005.
3. Okoli, C. A guide to conducting a standalone systematic literature review. *Commun. Assoc. Inf. Syst.* **2015**, *37*, 879–910.
4. Thomson Reuters. Web of Science 1 Billion Cited References and Counting. Thomson Reuters 2017. Available online: http://stateofinnovation.thomsonreuters.com/web-of-science-1-billion-cited-references-and-counting?utm_source=false&utm_medium=false&utm_campaign=false (accessed on 5 January 5 2017).
5. Benítez, I.; Quijano, A.; Díez, J.; Delgado, I. Dynamic clustering segmentation applied to load profiles of energy consumption from Spanish customers. *Int. J. Electr. Power Energy Syst.* **2014**, *55*, 437–448. [CrossRef]
6. Gouveia, J.P.; Seixas, J. Unraveling electricity consumption profiles in households through clusters: Combining smart meters and door-to-door surveys. *Energy Build.* **2016**, *116*, 666–676. [CrossRef]
7. Kavousian, A.; Rajagopal, R.; Fischer, M. Determinants of residential electricity consumption: Using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants' behavior. *Energy* **2013**, *55*, 184–194. [CrossRef]
8. Mcloughlin, F.; Duffy, A.; Conlon, M. Characterising domestic electricity consumption patterns by dwelling and occupant socio-economic variables: An Irish case study. *Energy Build.* **2012**, *48*, 240–248. [CrossRef]
9. Flath, C.; Nicolay, D.; Conte, T.; Van Dinther, C.; Filipova-Neumann, L. Cluster analysis of smart metering data: An implementation in practice. *Bus. Inf. Syst. Eng.* **2012**, *4*, 31–39. [CrossRef]
10. Chicco, G.; Napoli, R.; Piglione, F. Comparisons among clustering techniques for electricity customer classification. *IEEE Trans. Power Syst.* **2006**, *21*, 1–7. [CrossRef]

11. Mcloughlin, F.; Duffy, A.; Conlon, M. A clustering approach to domestic electricity load profile characterisation using smart metering data. *Appl. Energy* **2015**, *141*, 190–199. [CrossRef]
12. Ozawa, A.; Furusato, R.; Yoshida, Y. Determining the relationship between a household's lifestyle and its electricity consumption in Japan by analyzing measured electric load profiles. *Energy Build.* **2016**, *119*, 200–210. [CrossRef]
13. Granell, R.; Axon, C.J.; Wallom, D.C.H. Impacts of raw data temporal resolution using selected clustering methods on residential electricity load profiles. *IEEE Trans. Power Syst.* **2015**, *30*, 3217–3224. [CrossRef]
14. Al-otaibi, R.; Jin, N.; Wilcox, T.; Flach, P. Feature construction and calibration for clustering daily load curves from smart-meter data. *IEEE Trans. Ind. Inform.* **2016**, *12*, 645–654. [CrossRef]
15. Coke, G.; Tsao, M. Random effects mixture models for clustering electrical load series. *J. Time Ser. Anal.* **2010**, *31*, 451–464. [CrossRef]
16. Chicco, G.; Sumaili Akilimali, J. Renyi entropy-based classification of daily electrical load patterns. *IET Gener. Transm. Distrib.* **2010**, *4*, 736–745. [CrossRef]
17. Macedo, M.N.Q.; Galo, J.J.M.; De Almeida, L.A.L.; Lima, A.C.D.C. Demand side management using artificial neural networks in a smart grid environment. *Renew. Sustain. Energy Rev.* **2015**, *41*, 128–133. [CrossRef]
18. Albert, A.; Rajagopal, R. Smart meter driven segmentation: What your consumption says about you. *IEEE Trans. Power Syst.* **2013**, *28*, 4019–4030. [CrossRef]
19. Gulbinas, R.; Khosrowpour, A.; Taylor, J. Segmentation and classification of commercial building occupants by energy-use efficiency and predictability. *IEEE Trans. Smart Grid* **2015**, *6*, 1414–1424. [CrossRef]
20. Viegas, J.L.; Vieira, S.M.; Melício, R.; Mendes, V.M.F.; Sousa, J.M.C. Classification of new electricity customers based on surveys and smart metering data Commission for Energy Regulation. *Energy* **2016**, *107*, 804–817. [CrossRef]
21. Bishop, C.M. *Pattern Recognition and Machine Learning*, 1st ed.; Springer: New York, USA, 2007.
22. Räsänen, T.; Voukantsis, D.; Niska, H.; Karatzas, K.; Kolehmainen, M. Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data. *Appl. Energy* **2010**, *87*, 3538–3545. [CrossRef]
23. Park, S.; Ryu, S.; Choi, Y.; Kim, J.; Kim, H. Data-driven baseline estimation of residential buildings for demand response. *Energies* **2015**, *8*, 10239–10259. [CrossRef]
24. Friedman, J.; Hastie, T. *The Elements of Statistical Learning*, 1st ed.; Springer: New York, NY, USA, 2001.
25. Kang, J.; Lee, J. Electricity customer clustering following experts' principle for demand response applications. *Energies* **2015**, *8*, 12242–12265. [CrossRef]
26. López, J.J.; Aguado, J.A.; Martín, F.; Mu, F.; Rodríguez, A.; Ruiz, J.E. Hopfield—K-Means clustering algorithm: A proposal for the segmentation of electricity customers. *Electr. Power Syst. Res.* **2011**, *81*, 716–724. [CrossRef]
27. Tong, X.; Li, R.; Li, F.; Kang, C. Cross-domain feature selection and coding for household energy behavior. *Energy* **2016**, *107*, 9–16. [CrossRef]
28. Simshauser, P.; Downer, D. On the inequity of flat-rate electricity tariffs. *Energy J.* **2016**, *37*, 199–229. [CrossRef]
29. Haben, S.; Singleton, C.; Grindrod, P. Analysis and clustering of residential customers energy behavioral demand using smart meter data. *IEEE Trans. Smart Grid* **2015**, *7*, 136–144. [CrossRef]
30. Dansk, E. *Giv Energien Videre: Nye Energipolitiske Visioner Og Udfordringer 2020–2030*; Dansk Energi: Copenhagen, Denmark, 2015.
31. Governmental Offices of Sweden. Fossil Free Sweden N.D. Available online: http://www.government.se/government-policy/fossil-free-sweden/ (accessed on 27 January 2017).

Paper 2 - Electricity Consumption Clustering Using Smart Meter Data

energies

MDPI

Article

Electricity Consumption Clustering Using Smart Meter Data

Alexander Tureczek [1,*], Per Sieverts Nielsen [1] and Henrik Madsen [2]

[1] Systems Analysis, the Department of Management Engineering, Technical University of Denmark, 2800 Kgs. Lyngby 2800 Kgs, Denmark; pernn@dtu.dk
[2] Dynamical Systems, the Department Compute, Technical University of Denmark, 2800 Kgs. Lyngby 2800 Kgs, Denmark; hmad@dtu.dk
* Correspondence: atur@dtu.dk; Tel.: +45-2346-0989

Received: 23 February 2018; Accepted: 4 April 2018; Published: 6 April 2018

Abstract: Electricity smart meter consumption data is enabling utilities to analyze consumption information at unprecedented granularity. Much focus has been directed towards consumption clustering for diversifying tariffs; through modern clustering methods, cluster analyses have been performed. However, the clusters developed exhibit a large variation with resulting shadow clusters, making it impossible to truly identify the individual clusters. Using clearly defined dwelling types, this paper will present methods to improve clustering by harvesting inherent structure from the smart meter data. This paper clusters domestic electricity consumption using smart meter data from the Danish city of Esbjerg. Methods from time series analysis and wavelets are applied to enable the K-Means clustering method to account for autocorrelation in data and thereby improve the clustering performance. The results show the importance of data knowledge and we identify sub-clusters of consumption within the dwelling types and enable K-Means to produce satisfactory clustering by accounting for a temporal component. Furthermore our study shows that careful preprocessing of the data to account for intrinsic structure enables better clustering performance by the K-Means method.

Keywords: smart meter analysis; electricity consumption clustering; data analysis; K-Means; autocorrelation

1. Introduction

The number of days that Denmark fully covers its electricity demand through renewable sources is increasing. By the end of 2020, renewable electricity production in Denmark is projected to cover an average of 84% of electricity demand [1]. Though there is still a deficit of renewables in the system, the gap is closing, also at a European scale [2]. The caveat is that renewables induce volatility in the electricity grid as the production is tied to uncontrollable sources. A deeper understanding of electricity demand can help alleviate the implications of the volatile production, by promoting flexible consumption through tariff incentives.

The advent of residential electricity smart meters has enabled utilities to record and monitor electricity consumption by the minute. Recording electricity consumption at this unprecedented granularity can help us understand electricity demand in more detail. Analyzing consumption patterns can enable electricity utilities to develop targeted tariffs for individual groups mitigating production volatility by harnessing the flexibility of consumers.

The future electricity grid is expected to experience growing demand from the electrification of transportation [1] and the increased application of electric heat pumps. The introduction of renewable resources in the electricity sector therefore introduces significant challenges. The expected increase in demand and volatility in electricity production will put a strain on the entire distribution and

transmission grid. Demand flexibility has been discussed as a means to match demand with the volatility in production. To evaluate demand flexibility, a deeper understanding of consumption patterns is essential.

The application of smart meter data to cluster electricity consumption is a research field that has been gaining momentum over the past decade, beginning with [3], which analyzed smart meter electricity data, clustering methods and validation. In the electricity smart meter literature, K-Means is a very prevalent [4] method for clustering. The clusters created often exhibit variation to such an extent that clusters overlap, resulting in academically viable but practically indistinguishable clusters.

This paper will apply modern data mining techniques and methods from signal analysis to reduce the cluster overlap. The proposed methods will enable K-Means to analyze intrinsic data information which was previously ignored by the clustering. Reducing the overlap will produce more distinguishable and generally applicable cluster solutions. Data from more than 34,000 household electricity smart meters are included in the analysis performed in this paper. This paper contributes to the electricity smart meter literature through the following:

- Presenting a cluster analysis of Danish household electricity consumption data.
- Confirmation of autocorrelation in the data, information which K-Means is unable to incorporate in the clustering.
- Transformation and extraction of input data features enabling K-Means to account for autocorrelation in the clustering. This can easily be extended to include other data structures.
- Extending the concept of cross-validation to unsupervised learning employing cluster validation indices resulting in variability estimates of the resulting clustering performance.

The remainder of this paper is divided into six sections. First, Section 2 describes the current state of the art of smart meter electricity consumption classification, followed by a data summary and preprocessing in Section 3. Section 4 outlines the methodology applied in this paper. In Section 5 we apply the methodology to the smart meter electricity consumption data, followed by a discussion of the results in Section 6, and Section 7 concludes with the papers contributions.

2. Literature Review

This section presents a review of the current state of the art in smart meter electricity consumption clustering. The foundation for the study is [4], which conducts a systematic review of the current state of the art in smart meter data analytics. The paper evaluates approximately 2100 unique peer-reviewed papers and presents three main finding related to clustering methods, data and cluster validation.

Several methods for clustering have been applied and the most prevalent is K-means [5,6] and derivatives such as fuzzy K-Means [7,8] and adaptive K-Means [9]. Further algorithms like hierarchical clustering [10,11], and random effect mixture models [12,13] are also popular. Many of the papers apply K-Means for baseline clustering and compare more advanced methods to this baseline [14–16], with inconclusive outcomes regarding the best method for clustering. Some papers make an effort to preprocess the smart meter data; popular preprocessing methods are principal component analysis and factor analysis for dimensionality reduction [17,18] and self-organizing maps for 2 Dimensional representation of the data [3,10]. All identified methods are not particularly well-suited to time series data, such as smart meter data. Consequently, the clustering methods applied to the data do not leverage the intrinsic temporal data structure hidden in the smart meter data.

Many of the papers identified in [4] fail to acknowledge smart meter readings as time series data, a data type which contains a temporal component. Only one paper recognized the time series properties through the application of Fourier transformation, which maps data from the time to the frequency domain and subsequently applies K-Means to cluster by largest frequency [7]. The omission of the time series structure in the analysis leads to the application of methods that are not designed for handling temporal components. K-Means ignores autocorrelation, unless the input data is preprocessed; methods for preprocessing input data to enable K-Means to account for autocorrelation are described in [19].

In [20,21], principal component analysis and similarity measures for time series evaluation of generic data are discussed. The conclusions are applicable to smart meter data, although the method works best with fewer meters than recordings and thus, conversely, the dataset expands.

The clusters identified in the papers are validated by a variety of indices, with the most prevalent being the cluster dispersion index (CDI) [22–24], the Davies–Bouldin index (DBI) [25,26] and the mean index adequacy (MIA) [8,13].

This paper will describe methods for preprocessing smart meter data to enable K-Means to evaluate autocorrelation in data. These methods will make it possible to exploit hidden structures and thus increase the amount of information applicable for clustering.

3. Data Summary and Preparation

This section introduces the smart meter electricity consumption data that will be analyzed for the remainder of this paper. The data is kindly provided by SydEnergi, the largest electricity utility company in southern Denmark.

This paper analyzes consumption patterns for apartments and (semi)detached houses connected to the district heating system in the city of Esbjerg. It covers four postal codes—6700, 6705, 6710, and 6715—and the two selected household types are expected to behave identically. There were initially 34,000+ consumers of these two types in Esbjerg, each with a smart meter installed that records consumption every 60 min. We only analyzed these two residential categories as we were interested in analyzing consumption differences within consumer groups and not across different housing types.

The literature does not advise on the time length for analyzing consumption patterns. Paper [16] analyzes load profiles with a consumption window of one week, which is also the consumption window that we selected for this study. We selected the second week of January 2011, starting on Monday the 10th and ending Sunday the 16th, with both days included. With consumption recorded each 60 min, this yields 24 recordings per day for a total of 168 recordings per meter across the seven days.

The precise number and types of smart meter data employed in this paper is described in Table 1. The accompanying waterfall table (Table 2) illustrates the effect of the preprocessing on the final data set size.

Table 1. Initial data description of SydEnergi data for the city of Esbjerg, comprising 13 distinct quantitative measures of the data applied in the paper. As introduced in [4].

Data Description	Value
Country	Denmark
Region	Region Syd (Region South) postal codes: 6700, 6705, 6710, 6715 (City of Esbjerg)
Supplier	SydEnergi Electricity Utility
Initial Size	34,418 m
Clear Reduction	Confer Table 2.
Missing Values	70 m
Final Size	32,241
Recording Frequency	60 min
Start	10 January 2011
End	16 January 2011
Length	168 observations (hourly readings)
Type	Single family house (18,058 initial size) Apartments (15,721 initial size) both heated via district heating.
Referral	Data has never before been referenced.

Before analysis, the data was preprocessed to remove missing data and other undesirable traits. A two-stage process for cleaning the data was applied. Stage 1 involved a simple descriptive statistical examination of the data, ensuring the removal of; missing values, zero mean consumption, zero median consumption, and zero variance, all of which would indicate missing consumption information.

This preprocessing is outlined in a waterfall statistic, seen in Table 2, which presents the effect of each step of the preprocessing. For more advanced anomaly detection methods see [27]. Stage 2 exploited the fact that the data set encompassed data from the subsequent third week of January from the 17th–23rd. This helped us to identify meters that were behaving irregularly in week two, by evaluating the week-on-week consumption change. This change can be an indication of vacant dwellings with a subsequent consumption increase, e.g., returning from vacation. We defined irregular as a week-on-week consumption change of more than 200%. Meters that exhibited this consumption pattern were removed from the data set.

Table 2. Data cleaning waterfall. *Filter* indicates the removal criteria and *Meters* show the remaining meters after the application of the filter. *Discard* is the number of meters discarded through the filtering. *Final bulk* is the number of meters ready for analysis after the cleaning of data.

Filter	Meters	Discard	Note
Initial Data	34,418	-	Original data
Missing	34,348	70	Removal of meters with missing recordings
Mean Zero	33,325	1023	Removal of meters with 0 mean indicating no consumption
Median Zero	32,745	580	Removal of meters with 0 median indicating no consumption
Variance Zero	32,745	0	Removal of meters with 0 variance indicating flat consumption
Consumption < 0	32,744	1	Removal of meters with <0 consumption indicating prosumers
Overlapping	32,586	158	Overlapping with 2nd week for comparison
+200% Increase	32,241	345	+200% consumption increase from (10th–16th) to (17th–23rd)
Final bulk	32,241	-	Final number of meters included in analysis

Figure 1 shows four different meters that exhibit week-on-week consumption changes above 200% percent. In the figure, the consumption change indicates a return to the dwelling; we were not interested in clustering vacant dwelling consumption and accordingly removed meters with a 200% increase in consumption.

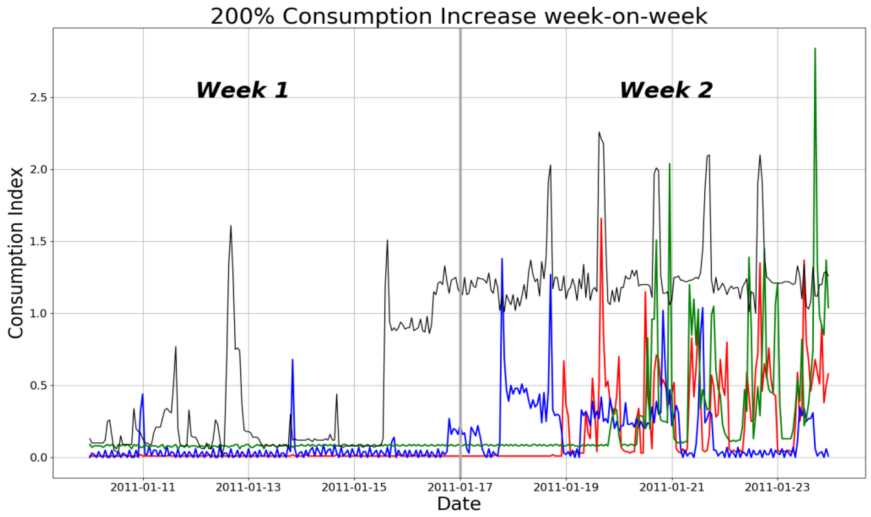


Figure 1. Four different meters all exhibiting a week-on-week consumption increase above 200%. This is to filter out dwellings that were vacant during the week analyzed, as we were not interested in clustering standby consumption.

4. Methodology

This section describes the theoretical statistical framework that we applied to analyze the smart meter data. Section 4.1 starts with a discussion of the concept of statistical learning and presents a flow chart illustrating the process applied in this paper. The literature review in Section 2 identified

K-Means clustering as the most prevalent clustering method for electricity smart meter consumption data. Section 4.2 discusses the K-Means clustering method and the importance of normalization. Section 4.3 includes a discussion of cluster validation with the subsequent description of four selected indices—MIA, cluster dispersion index, the Davies–Bouldin index and the silhouette index. This section also includes a description of the unsupervised cross-validation applied in this paper. Sections 4.4 and 4.5 discuss autocorrelation feature extraction and wavelet transformation, which are methods that can enable K-Means to include temporal components in the clustering process.

4.1. Statistical Learning

The statistical segmentation of data into smaller more homogeneous subsets is carried out by applying supervised or unsupervised learning. The distinction between supervised and unsupervised learning is bound to differences in the initial problem conditions. For supervised learning problems there exist some known class labels and knowledge of the membership attributes of a class. This membership knowledge is used to create a mathematical function that maps the observations to classes. For unsupervised learning, class labels do not exist. In unsupervised learning there exists no apparent external or internal information that can unambiguously identify the potential underlying clusters. Different methods have been developed in an effort to remedy the problem and enable unsupervised clustering, but the clusters identified in this way are rarely stable and unique. There exist several techniques for unsupervised clustering; popular methods include K-means and hierarchical clustering.

This paper introduces the extraction of data features to enable K-Means to account for the temporal component in smart meter data. Three different manipulations of the input data were investigated—normalization, wavelet transformation and autocorrelation feature extraction. Figure 2 illustrates a process overview, where blue boxes indicate the processes that all methods were subjected to. All analysis was drawn from the *data*, *preparation* of data and *clustering*. This paper introduces three different data manipulation methods prior to clustering, to enable K-Means to account for intrinsic information. The methods applied were autocorrelation feature extraction (red), normalization (black), and wavelets (green). We applied normalization to the wavelet transformation before clustering.

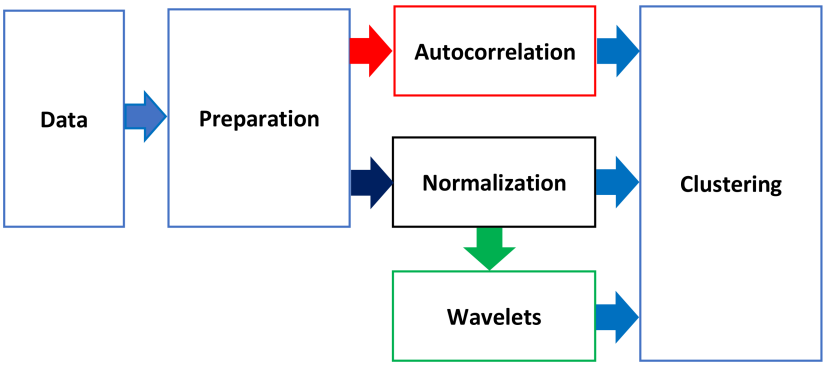


Figure 2. Methodology flow chart. This chart illustrates the different data processing methods applied. The (blue) boxes indicate processes to which all methods were applied, namely *data*, *preparation* and *clustering*. After preparation, autocorrelation (red) indicates the extraction of autocorrelation features. Normalization (black) was applied both as a sole processing method, but also in preparation for wavelet transformation (green).

Smart meter data is recorded over time; accounting for the temporal component, which can convey information about the data patterns. By default, K-Means clustering does not consider this temporal component. Thus, a very important feature of the data—the temporal component—is not employed in the clustering. Figure 3 shows data with and without a temporal component; the left side shows data where the temporal component has been collapsed. It is not possible to estimate whether the data

overlaps or is just very close in distance. The right side shows the exact same data with the temporal component reinstated. From the right side it can clearly be seen that there is a temporal structure in the data, this component reveals three different non-overlapping cosine structures. The temporal component accounts for intrinsic data information that the K-means and other unsupervised methods do not evaluate when clustering. This paper will present methods to alleviate the problem and enable the K-Means method to account for temporal structures. Preprocessing the data before clustering with K-Means can help to include the structure from the right side of Figure 3.

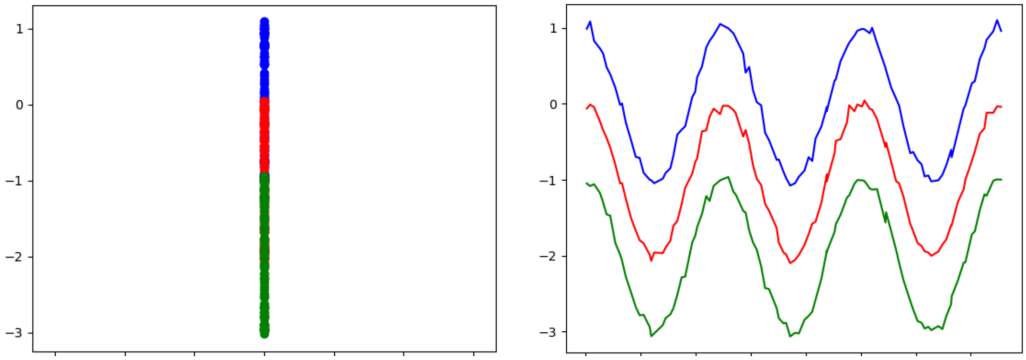


Figure 3. (**Left**): A scatter of points collapsed to have no temporal component. The three colors indicate three different clusters, but it is not possible to identify overlap. (**Right**): The scatter has been expanded by its original temporal component.

In electricity smart meter data analytics, the clustering methodology is often either K-Means or hierarchical clustering. These are simple and robust methods that perform reasonably well under various circumstances. Both employ a distance measure for clustering, and the selection of distance measure can heavily influence the shape of the clusters [28].

In the absence of knowledge of the true clusters, and to avoid the trivial clustering case of assigning one cluster to each observation, reducing the variability to 0, several cluster validation indices were introduced. These indices evaluated the intra-cluster distance and related it to the inter-cluster distance. Often the indices favor a clustering solution that minimizes the intra-cluster distance while maximizing the inter-cluster distance.

4.2. K-Means

As described in [4], the K-Means method is the most prevalent technique for electricity smart meter consumption clustering. K-Means is a simple and robust algorithm for partitioning n observations into k clusters. This is done by assigning each of the n observations to the closest cluster centroid given some distance measure. Due to random initialization of the K-Means algorithm, it can result in locally optimal solutions. It is advised to rerun the clustering several times with different initial random seeds and select the clustering that yields the best discriminatory performance [29].

The K-Means implementation employed in this paper is the SKlearn data analysis package for Python version 0.18.2 [30]. We used the SKlearn default settings for maximum iterations until convergence (max_iter) 300. The K-Means was by default randomly initialized 10 times. The initial random seed for testing purposes in this paper was set to 12345.

Even though K-Means yields robust solution it is important to recognize that K-Means only evaluates data from a distance perspective, which from a smart meter data perspective implies that each time step is evaluated independently without correlation to the neighboring time steps. That is; K-Means evaluates all meter readings at $t = 0$ without regards to any structure or correlation effect with neighboring time steps such as $t = 1$. Especially with time series, autocorrelation is an integral aspect, and for electricity consumption we expect there to be some recurrent structure in the

consumption patterns. K-Means does not evaluate this structure, however, through feature extraction it is possible to account for the autocorrelation in the input data [19], enabling K-Means to include this information in the clustering. This paper applies both autocorrelation feature extraction and the wavelet transformation described in Sections 4.4 and 4.5 to account for the autocorrelation.

Normalizing the smart meter time series makes the data fit the interval [0–1]; in [31], normalization was applied to smart meter data. This process makes it possible to identify time series with equivalent consumption patterns instead of identical consumption volumes. As the focus of this paper is clustering by consumption pattern, we normalize by:

$$Normalization = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

4.3. Cluster Validation

In unsupervised learning there exists no natural quantification of the discrepancy between model and truth, as the true clusters are unknown. The need for evaluating the performance of unsupervised methods has resulted in the development of various cluster evaluation indices [13]. This paper has, based on the prevalence found in [4], selected four prominent indices for validation, namely *MIA*, the *cluster dispersion index* (*CDI*), *Davies–Bouldin index* (*DBI*) and the *silhouette* index. The indices each evaluate different properties of the clusters. Even though none of the indices can identify the true underlying structure, their values for different number of clusters can give an indication of how many clusters to retain in the final clustering. Plotting the progression of the indices as a function of clusters allows for visual inspection, where abrupt changes in their decline or fluctuating pattern can help select the number of clusters within a given data set [19]. We advise the evaluation of several indices jointly, as the combination can be applied to strengthen the argument for the selection of a specific number of clusters.

4.3.1. Mean Adequacy Index (*MIA*)

The *MIA* index calculates the square root of the average distance from each member of a class to the class centroid and scales it by the number of classes *K*.

$$MIA = \sqrt{\frac{1}{K}\sum_{k=1}^{K} d^2(C_k)}, \tag{2}$$

where $d^2(C_k)$ is the squared average distance within cluster k. The *MIA* index is a measure of within-class dispersion. Large distances within the class indicate a poor fit; high index values indicate large within-cluster dispersion.

4.3.2. Cluster Dispersion Index (*CDI*)

The *CDI* is a revised version the MIA index scaled by the average cluster distance $d(C)$. The *CDI* prefers large inter-cluster distances and small intra-cluster distances [24].

$$CDI = \frac{1}{d(C)}\sqrt{\frac{1}{K}\sum_{k=1}^{K} d^2(C_k)} \tag{3}$$

Smaller values indicate better clustering. $d(C)$ is the average cluster distance between any two clusters in the clustering, while $d^2(C_k)$ is the average squared within the cluster distance.

4.3.3. Davies–Bouldin Index (*DBI*)

The *DBI* evaluates the overlap between clusters. This is done by evaluating the average intra-cluster distance, given by $diam(C_i)$, of all clusters i and subsequently comparing all pairs of clusters divided by their centroid distance $d(C_i, C_j)$ and selecting the maximum distance for each class.

$$DBI = \frac{1}{K}\sum_{i=1}^{K} \max_{j \neq i} \frac{diam(C_i) + diam(C_j)}{d(C_i, C_j)} \tag{4}$$

Smaller values of *DBI* implies that the K-means clustering algorithm separates the data set properly [11].

4.3.4. Silhouette Index

The *Silhouette* index evaluates $C(x)$ the average distance between each vector x within a class C. While $C'(x)$ is the minimum distance from a vector in class C to a vector not in C, scaled by the maximum distance between two classes C *and* C' [4].

$$Silhouette = \frac{c'(x) - c(x)}{\max\{c(x), c'(x)\}} \tag{5}$$

$$c'(x) = \min_{y \in C'} d(x, y) \tag{6}$$

The index is bound in the interval $[-1, 1]$, where higher values are better; negative values indicate misclustering [31].

4.3.5. Unsupervised Cross-Validation

Cross-validation is an effort to increase model robustness by dividing the data set into a training and a test set. The training set is used to train the model and the test set is used test the model on an "unknown" data set. The process helps quantify model stability and helps reduce the chance of overfit. For cross-validation to achieve its purpose of reducing overfit and evaluating model performance, there needs to exist a measure of fit. For unsupervised learning no such fit exists [32]; to remedy this situation we regard the cluster validation indices as the measure of fit, creating a pseudo cross-validation measure for the fit of our clustering. This pseudo cross-validation enables us at each number of clusters to evaluate the maximum and minimum value of the index and thus how stable the index is for a given number of clusters. This paper applied 10-fold cross validation to the indices.

4.4. Autocorrelation Feature Extraction

In time series analysis autocorrelation is an essential concept, encompassing the temporal component of the data, e.g., the time dependency in a data series. Autocorrelation is, like correlation, a standardized version of covariance, and is calculated like correlation but as a function of time steps. It quantifies the relation between time steps, called lags. Plotting the autocorrelation coefficients as a function of lag reveals important structures of the data such as trends, seasonality and the stability of the time series [33]. Figures 4–6 show different consumption recordings, illustrating different consumption patterns and different autocorrelation functions, with 48 lags. The left side shows the original consumption, while the middle shows the autocorrelation coefficient (solid line) and the 95% confidence interval (dashed line). The right side of the figures shows the significant autocorrelation coefficients. The figures illustrate differences in autocorrelation structures. In Figure 4 the lags indicate no daily cycle and only immediate lags are significant. Figures 5 and 6 exhibit a periodicity in the autocorrelation near lag 20, indicating a recurrent pattern for both consumers.

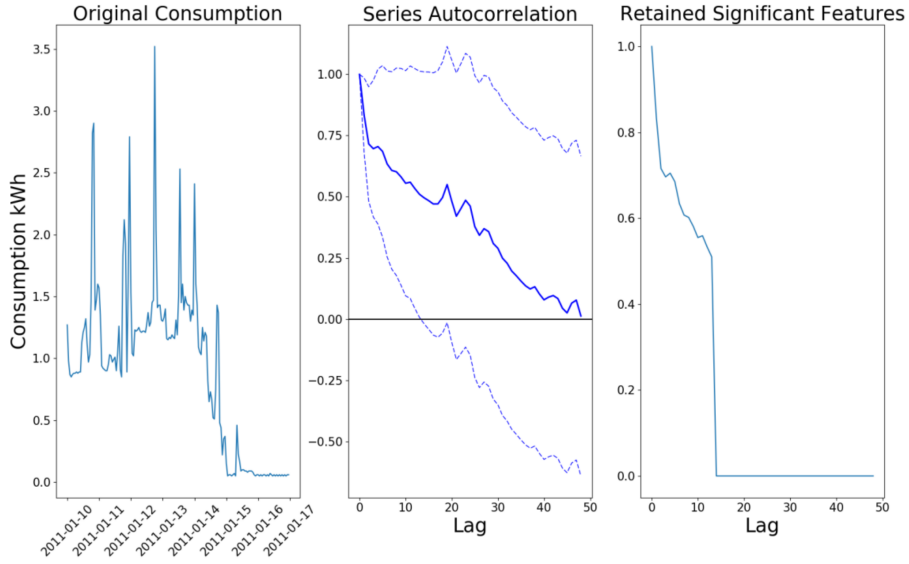


Figure 4. (**Left**) The original consumption profile, (**Middle**) the solid line is the autocorrelation coefficient, dashed lines are the 95% confidence intervals. (**Right**) Significant autocorrelation coefficients retained as meter features and applied as input to K-Means clustering. The significant lags only include the first 14 lags, indicating no recurrent pattern.

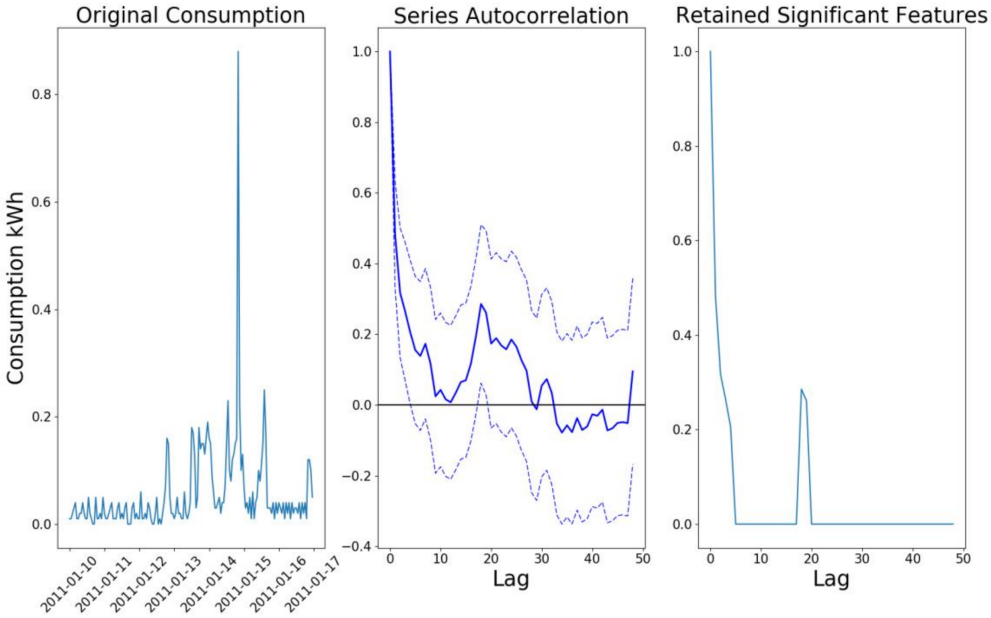


Figure 5. (**Left**) The original consumption profile, (**Middle**) the solid line is the autocorrelation coefficient, dashed lines are the 95% confidence intervals. (**Right**) Significant autocorrelation coefficients retained as meter features and applied as input to K-Means clustering. The significant lags include lags from the first five lags and a recurrence at around lag 20, indicating some periodicity in the consumption.

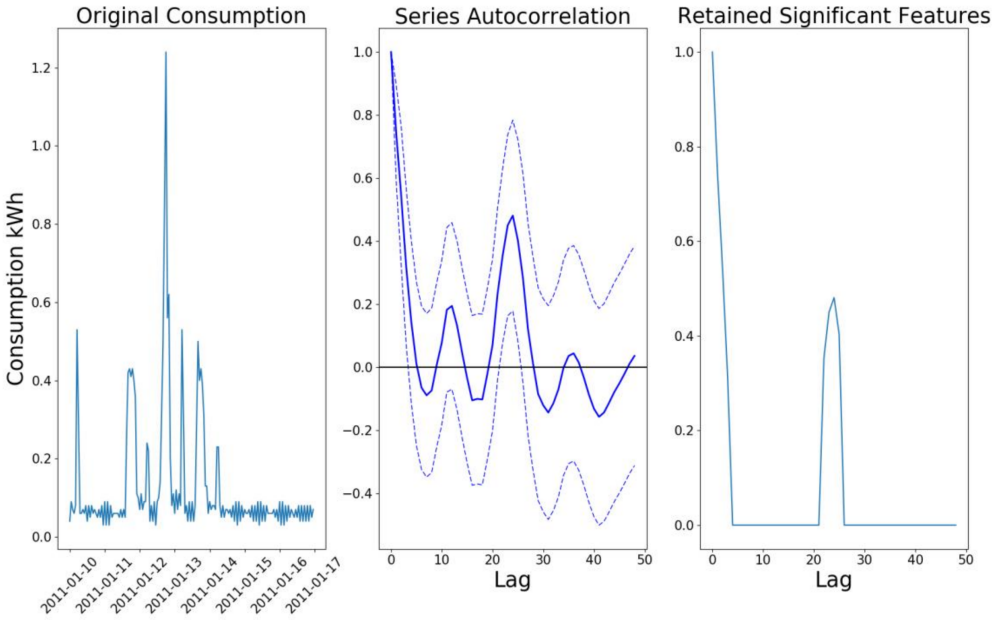


Figure 6. (**Left**) The original consumption profile, (**Middle**) the solid line is the autocorrelation coefficient, dashed lines are the 95% confidence intervals. (**Right**) Significant autocorrelation coefficients retained as meter features and applied as input to K-Means clustering. The significant lags include lags from the first five lags and a recurrence at around lag 20, indicating some periodicity in the consumption. The significant lags are distinct from the lags in Figure 5.

4.5. Wavelet Feature Extraction

The wavelet transformation is a basis transformation using wavelet basis functions; wavelets can represent smooth and locally non-smooth functions. Wavelets have time and frequency localization, effectively linking time and frequency in contrast to the Fourier transformation which only allows frequency localization [29]. Wavelets are especially well suited for analyzing high frequency data because of their ability to capture global smoothness and local spikes in the signal [34], while filtering out high frequency noise [35]. The application of wavelets for time series feature extraction in this study was inspired by [36]. In the process of filtering high frequency data, wavelets perform efficient data compression by removing non-significant coefficients. Often this process removes a considerable number of coefficients. The decomposition of the signal into wavelet coefficients is not easily interpretable by humans but are readily applicable as input for the K-Means algorithm. The wavelet coefficients are uncorrelated [37].

Choosing a suitable wavelet is difficult, as the scaled basis wavelet must be able to encompass the structure of the original signal. We applied the Coiflet 8 wavelet seen in Figure 7, which is highly fluctuating, enabling the encapsulation of high frequency data. We removed non-significant coefficients by applying universal thresholding [38] to the wavelet coefficients. The Python wavelet package PyWt [39] was utilized for the wavelet analysis performed in this paper.

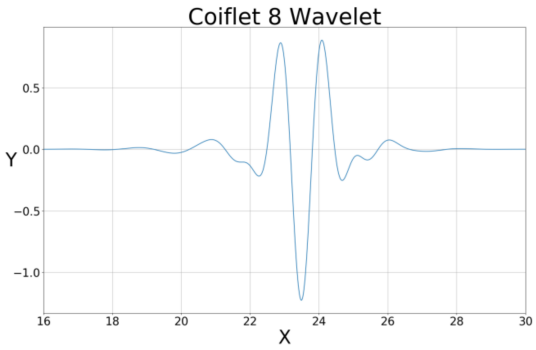


Figure 7. The Coiflet 8 wavelet applied in this paper. It exhibits a similar structure to the meter data, making it a suitable candidate for the wavelet transformation of the input data.

5. Results

This section will describe the results obtained by applying the methodology introduced in Section 4 to the dataset described in Section 3. Section 5.1 will describe clustering with normalized data, while Section 5.2 describes the application of the wavelet transformation and Section 5.3 describes the influence of the autocorrelation feature extraction on the clustering performance. Finally, Section 5.4 will summarize the performed clustering solutions.

5.1. Cluster Permance: Normalized Data

In [19] and various other papers, clustering smart meter consumption by only normalizing data has produced acceptable clustering performance. Clustering the SydEnergi data by only normalizing the data produced the inconclusive cluster validation index graphs seen in Figure 8, indicating a lack of identifiable clusters. Figure 8 shows how the cluster validation indices develop as function of the number of clusters. The dashed lines surrounding the individual indices indicate the maximum and minimum observed values at each selection of clusters calculated by pseudo cross-validation, as described in Section 4.3.5.

As described in Section 4.3, we were looking for an elbow break in the index development, indicating that more clusters will not improve the clustering. The silhouette and MIA indices exhibited very small changes, indicating stability, and both flattened almost immediately, as they were questionable with regards to their performance on the SydEnergi data. Arguably, the silhouette index indicates three clusters, but the structure was poorly defined in the graph and hence we discarded it as a possible optimum number of clusters. The inability of the K-Means to cluster the normalized data can in this case be attributed to the close resemblance of the households included in the study. We included only houses and apartments from the city of Esbjerg connected to district heating.

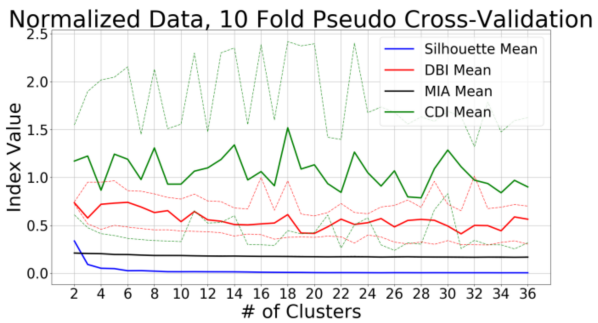


Figure 8. Cluster validation indicator development as a function of clusters. The silhouette and MIA validation indices were non-informative in this data set. While the CDI index exhibits large variation, there is no indication of optimum cluster selection, which was also the case for the DBI index.

Normalizing the SydEnergi input data prior to the K-Means clusters, the households included in this analysis were similar in overall grouping, but for 32,000+ consumers the method was expected to reveal sub-clusters. This is an indication that normalization of smart meter data in this case was so subtle that K-Means was unable to identify clusters.

5.2. Cluster Performance: Wavelet Transformation

The application of the wavelet transformation of input data resolves the autocorrelation, as the coefficients are uncorrelated. In effect, wavelet transformation performs dimension reduction, keeping the structure of the time series with a reduced number of coefficients. This makes the feature space very similar to the original space, thus—as seen in Figure 9—creating very similar cluster validation indices. In this case, the wavelets did not create more insightful index development and no apparent optimum number of clusters was identifiable. As the wavelet transform compressed the data, the inability to identify clusters was no surprise, as the compressed data was similar to the normalized data. The Python wavelet package; PyWt [39] was unable to calculate the silhouette due to memory overflow issues, attributable to the large data set, and this is thus not included.

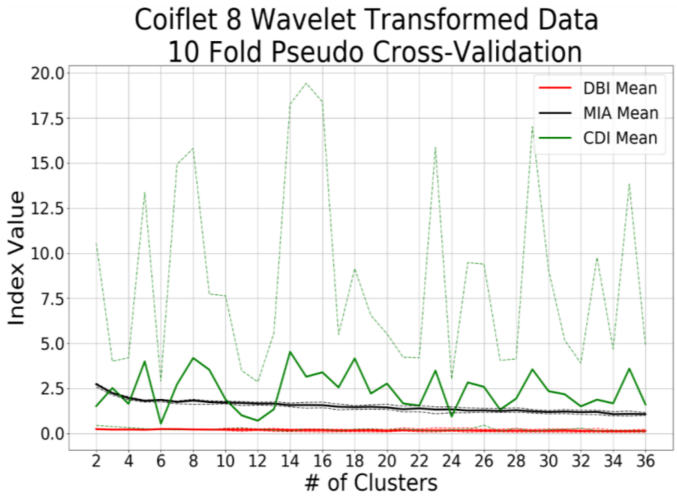


Figure 9. Cluster validation index development from 2 to 36 clusters using Coiflet 8 wavelet transformation. Significance was established applying universal thresholding. No apparent structure was found in the development of the three indices. As with normalized data, the CDI exhibited large fluctuations, while MIA and DBI had very controlled fluctuations.

5.3. Cluster Performance: Autocorrelation Feature Extraction

The autocorrelation feature extraction (ACF) method—described in Section 4.4—was applied to the data with 24 lags, equivalent to 24 h temporal information. Only statistically significant lags were retained as input data to the K-Means. The transformation reduced the data set size from 32,241 smart meters X 168 hours to 32,241 smart meters X 24 lags (hours). This is a clear reduction of the dataset, with a tangible effect on the computational cost of the K-Means clustering.

As with the normalized clustering, we calculated the cluster validation index for each number of clusters from 2 to 36; Figure 10 shows the index development. The solid line represents the average index value, with the corresponding dashed lines indicating the maximum and minimum observed values for any given cluster number. In Figure 10, the DBI index shows an "elbow break" at 12 clusters, combined with narrow minimum and maximum bands, implying that 12 clusters is optimum. The MIA and silhouette indices are almost horizontal throughout the entire span of clusters, with very small variation, giving no indication of cluster selection. In contrast, the CDI index exhibits large variation and a jagged horizontal development, also indicating no specific number of clusters. This indicates

that the autocorrelation features are potent for the identification of subtle differences in a perceived homogeneous group, enabling even finer clustering.

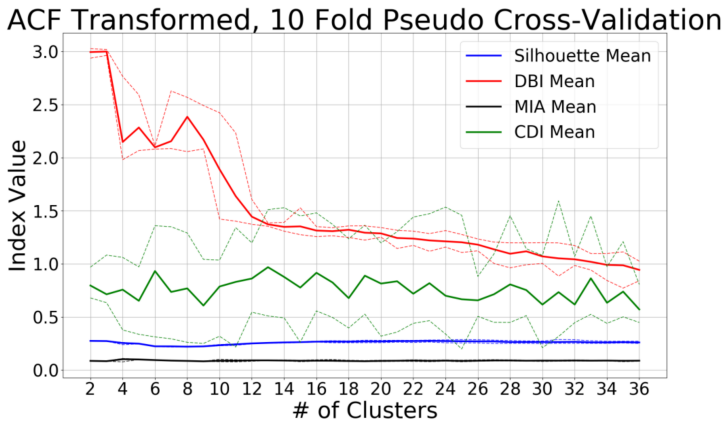


Figure 10. Cluster validation index development for the autocorrelation features (ACF). The DBI index shows a distinct "elbow" break at 12 clusters.

The corresponding plot of the different cluster means in the 12-cluster case is shown in Figure 11. Many of the clusters exhibit similar auto-correlation structures with slight variations in the value and lag offset. Generally, except for clusters 4 and 8, there is a short-term dependency of the past five lags (hours) with zero significant lags in the interval from 5 to 20 lags (hours), and then an indication of recurrent structure. Clusters 4 and 8 are distinctively different from all other clusters; cluster 4 shows a close to linear declining lag function, but no recurrent component, indicating no daily cycle in the consumption pattern. Cluster 8 also exhibits a close to linear decline throughout the 24 lags—except for some fluctuation in the very first lag—and no indication of a 24 hour trend. The remaining clusters have significant lags for the first and final five lags, with different offsets around lag 20 indicating a recurrent consumption pattern. The 12 clusters indicate similar consumption but with slight differences, these differences are attributable to diversity in consumption; although the overall consumption is similar, the finer details are amplified using the autocorrelation features.

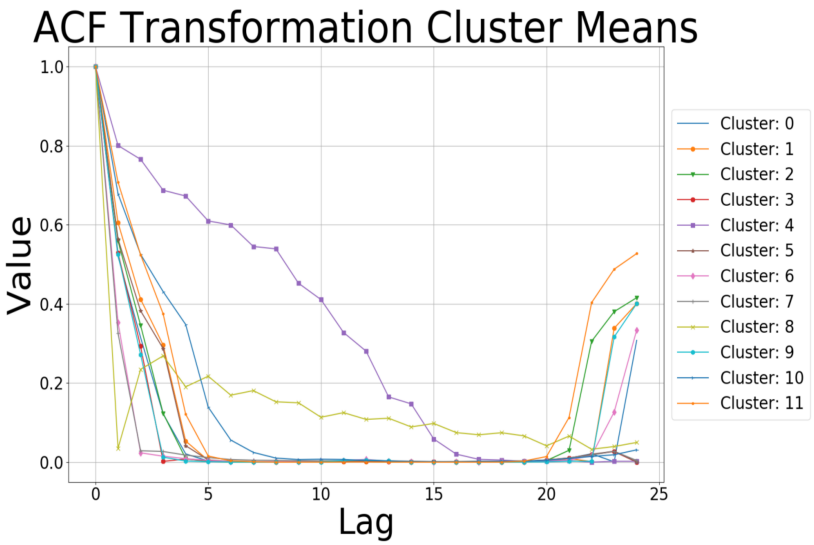


Figure 11. Plot of the 12 autocorrelation function cluster means (ACF) identified using the CDI. Clusters 4 and 8 are distinctly different showing linear decline and no recurrence. The remaining clusters exhibit a largely similar structure, with different values, and a different lag for recurrence.

The cluster composition of the 12 identified clusters gives further indication of the clustering performance. Table 3 presents an overview of each cluster's composition in terms of size, dwelling type distribution and postal code distribution. The clusters are well balanced, each accounting for approximately 10% of the total data. In each cluster there is a 40–60% penetration of apartments, indicating, as stated in Section 3, that electricity consumption is influenced more by inhabitants than by dwelling type; apartment or house. Finally, the distribution across postal codes is even according to the size of each postal code. This indicates overall balanced clusters and not just a clustering of select outliers, demonstrating no geographical clustering.

Table 3. Cluster composition table of the 12 different clusters. Only clusters 4 and 8 are markedly different from rest, with a very small cluster size. The remaining clusters sizes are well-balanced across all parameters.

Cluster Composition			Dwelling Type		Postal Code in Esbjerg			
Cluster	**Size**	**% of Total Data**	**Apartments**	**Houses**	**6700**	**6705**	**6710**	**6715**
0	3198	9.92%	1244	1954	1396	571	754	477
1	2456	7.62%	851	1605	976	460	609	411
2	3342	10.37%	1240	2102	1427	603	798	514
3	3988	12.37%	1920	2068	1953	739	763	533
4	239	0.74%	117	122	127	36	45	31
5	4295	13.32%	1854	2441	1956	846	888	605
6	3014	9.35%	1616	1398	1522	586	489	417
7	3590	11.13%	2237	1353	1976	674	539	401
8	405	1.26%	300	105	256	63	46	40
9	3703	11.48%	1476	2227	1568	670	868	597
10	1794	5.56%	859	935	875	344	347	228
11	2217	6.88%	946	1271	940	462	488	327
Total	32,241	100.00%	14,660	17,581	14,972	6054	6634	4581

5.4. Comparison of Results

The three different preprocessing methods of the K-Means input data yield very different results. In the data from Esbjerg, where the two household composition groups chosen are very similar, the normalization and the normalization + wavelet transformation were unable to provide any meaningful clustering solution for the data. There was no significant difference in the data structure between normalizing and wavelet transformation of the SydEnergi data. The wavelets do compress and remove autocorrelation, but do not provide the K-Means with the possibility to leverage the autocorrelation. For data where the normalization produced viable clustering solutions, the wavelet transformation was expected to do the same, but with a reduced number of dimensions and thus a significant reduction in computational effort.

With the SydEnergi data, the autocorrelation feature (ACF) method provided clustering solutions that leveraged the autocorrelation inherent in the data. This produced balanced clusters that encompassed the underlying structure found in the consumption patterns of the individual smart meters. The clustering solution generated by ACF was different from the normalization and wavelet transformation solution in that it provided more clusters, but also a different number of clusters. This difference was also observed in [19].

A measure for evaluating the clustering is analyzing the computational effort needed to perform the clustering. All three cases preprocessed the input data. The processes can be run in constant time and their influence on the overall runtime is negligible compared to K-Means lower bound runtime of $k\sqrt{n}$ [40] and upper bound of $O(k^n)$ [40], where k is clusters and n is observations. The reduction of the input data via the autocorrelation features or wavelet transformation can result in a significant decrease in the minimum and worst case computational effort needed to cluster the data [19], see Table 4.

Table 4. Clustering method runtime comparison. The normalized and wavelet methods were unable to provide meaningful clusters and are for comparison set to 12 clusters and 25% compression for wavelets. The autocorrelation and wavelet method reduced the dataset size, with a significant impact on the runtime. The table is an adaptation from a table in [19].

Processing	Normalization	Autocorrelation Features	Wavelet
Scaling/Transform	Constant time	Constant time	Constant time
Size of input data (n)	168 × 32k+	24 × 32k+	42 × 32k+
Best case running time (12 clusters)	$12^{\sqrt{168}}$	$12^{\sqrt{24}}$	$12^{\sqrt{42}}$
Worst case running time (12 clusters)	12^{168}	12^{24}	12^{42}

6. Discussion

The K-Means clustering algorithm is a simple, efficient and robust method for unsupervised clustering. It is readily implemented in many software suites and easy to apply to numeric data sets. However, the straightforward application of modern data mining software exposes possible pitfalls in data analysis. As this paper shows, it is not possible to calculate meaningful clusters from the SydEnergi data applying the K-Means method directly. Only through careful preprocessing of the input data to enable K-Means to account for temporal components did we calculate meaningful clusters. This demonstrates the importance of understanding and recognizing the data type under analysis. Failing to regard smart meter data as data evolving over time impairs the analysis by not encompassing all available information. Intrinsic data information is not generally utilized in smart meter analysis; reference [4] showed that data type knowledge was not consistently applied in the literature. In the case of smart meter data, the missing information is the autocorrelation, which quantifies how past observations influence current observations.

As described in Section 4.2, K-Means is unable to include autocorrelation, and in effect ignores this intrinsic information. In a supervised—e.g. regression—setting, this could potentially result in singularities, making the problem unsolvable, or at least rendering the coefficients unstable. K-Means robustness ignores this and creates a clustering regardless, not requiring the analyst to reflect upon model and data decisions. This paper has improved on the clustering performance of K-Means by enabling the algorithm to account for intrinsic information. This has been achieved through transformation and feature extraction based on data insights. The preprocessing of the input data enables K-Means to cluster data structures it was not originally intended to include. K-Means input preprocessing was successfully applied in this paper, but also to district heating data [19], where it was applied to reduce within-cluster variance.

K-Means is useful for prototyping, with extensive applications in smart meter clustering. However, the within-cluster variability is consistently large, such that the clusters overlap, delivering academically viable clusters with inconsequential practical value. The overlap results in indistinguishable clusters. There exists a gap in the literature on time series comparison, not just regarding clustering, but also the subsequent evaluation of the similarity of the time series. There exists some literature where various features are extracted from the individual time series and compared. This is a computationally expensive process, which is not always easily automatable. In general, the features proposed and traditional time series analysis have not yet been combined into a strong framework for comparing time series data. Ultimately there is a need for future research into statistically sound methods for evaluating the differences between time series, enabling researchers to better evaluate the resulting clusters and conclude on their (dis)similarity. Without better tools to evaluate differences in time series and reduce the within-cluster variability, smart meter consumption clustering could potentially linger as an academic exercise. This applies not only to smart meter data, but to time series clustering in general.

Not all transformations improve the clustering performance of K-Means. The paper applied several, with only successful application of autocorrelation features. Further, we conducted a principal component analysis with an ensuing substantial reduction in dimensions, however the subsequent

clustering of the transformed data showed no improvement compared to the normalization of the input data.

Wavelet transformation was applied and retained much of the general structure of the data in compressed form; thus the cluster validation index development closely resembled the development of the normalized data. The wavelet transformation removed autocorrelation and compressed the data by large factors, resulting in faster performance of the clustering, but with a similar result as that for the original uncompressed data.

The feature extraction methods applied in this paper also reduced the dimensionality of the input data set. This reduction had a significant impact on the computational cost of clustering smart meter data. The wavelet method compresses but maintains the original structure, enabling faster but similar clusters than normalized clustering, while the autocorrelation clusters around data features from the time series and produces different and—for the SydEnergi data—finer-grained clusters.

Mathematics provides a myriad of methods for data manipulation, which can help draw out intrinsic information from data. It requires that the analyst bring knowledge of the data and reflect upon the methods applicable, beyond the popular choices, and that they apply their knowledge to improve the model performance. This paper has shown that careful preprocessing of the data before clustering can improve the clustering performance in several ways, namely speed, the information included in clustering and better cluster definitions by measure of variance.

7. Conclusions

This paper has shown the existence of autocorrelation in specific smart meter electricity data. It is not a general proof of the existence of autocorrelation in all smart meter datasets, but is an indication that smart meter data needs to be examined for autocorrelation before analysis commences. This paper successfully extracted significant autocorrelation coefficients and incorporated them into subsequent clustering using K-Means.

The autocorrelation coefficients, regarded as features, enabled the K-Means algorithms to encompass autocorrelation and deliver more detailed clusters. The resulting clusters are well balanced, with an even distribution of dwelling type within each cluster and across different postal codes. Two clusters were distinctly different from the rest in their overall consumption profile but also in their size, being markedly smaller. In contrast, normalizing the smart meter electricity consumption data was unsuccessful in providing unambiguous clusters. Wavelet transformation of the input data to the K-Means was successful in compressing the data and removing multi-collinearity, but it did not succeed in identifying an optimum number of clusters. Furthermore, this paper implemented an unsupervised version of cross-validation enabling stability measures of the validation indices.

In conclusion, this paper has shown that the clever transformation of data prior to K-Means clustering can improve performance and enable K-Means to handle data and information of types for which it was not originally intended. This result makes it possible to produce clusters from smart meter data that are better defined through smaller clusters with less within-cluster variance.

Acknowledgments: This work is part of the CITIES project funded in part by the Danish Innovation found. Grant DSF 1305-00027B. The data was provided by SydEnergi.

Author Contributions: Alexander Tureczek conceived and designed the study with subsequent analysis. Alexander Tureczek and Per Sieverts Nielsen wrote the paper. Per Sieverts Nielsen and Henrik Madsen reviewed and proofread the paper. Henrik Madsen facilitated the data acquisition for this paper.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Dansk Energi (Danish Energy Association). *Giv Energien Videre—Nye Energipolitiske Visioner og Udfordringer 2020–2030*; Copenhagen, Denmark, 2015. Available online: https://www.danskenergi.dk/udgivelser/nye-energipolitiske-visioner-udfordringer-2020-2030-giv-energien-videre (accessed on 22 February 2018).
2. Eurelectric. *Vision for the European Electricity Industry*; 2018; p. 2. Available online: http://www.eurelectric.org/media/340222/vision_for_the_european_electricity_industry-2017-030-0781-01-e.pdf (accessed on 22 February 2018).
3. Chicco, G.; Napoli, R.; Piglione, F.; Postolache, P.; Scutariu, M.; Toader, C. Load pattern-based classification of electricity customers. *IEEE Trans. Power Syst.* **2004**, *19*, 1232–1239. [CrossRef]
4. Tureczek, A.M.; Nielsen, P.S. Structured Literature Review of Electricity Consumption Classification Using Smart Meter Data. *Energies* **2017**, *10*, 584. [CrossRef]
5. Rhodes, J.D.; Cole, W.J.; Upshaw, C.R.; Edgar, T.F.; Webber, M.E. Clustering analysis of residential electricity demand profiles. *Appl. Energy* **2014**, *135*, 461–471. [CrossRef]
6. Park, S.; Ryu, S.; Choi, Y.; Kim, J.; Kim, H. Data-Driven Baseline Estimation of Residential Buildings for Demand Response. *Energies* **2015**, *8*, 10239–10259. [CrossRef]
7. Ozawa, A.; Furusato, R.; Yoshida, Y. Determining the relationship between a household's lifestyle and its electricity consumption in Japan by analyzing measured electric load profiles. *Energy Build.* **2016**, *119*, 200–210. [CrossRef]
8. Tsekouras, G.J.; Hatziargyriou, N.D.; Member, S.; Dialynas, E.N.; Member, S. Two-Stage Pattern Recognition of Load Curves for Classification of Electricity Customers. *IEEE Trans. Power Syst.* **2007**, *22*, 1120–1128. [CrossRef]
9. Kwac, J.; Flora, J.; Rajagopal, R. Household energy consumption segmentation using hourly data. *IEEE Trans. Smart Grid* **2014**, *5*, 420–430. [CrossRef]
10. Räsänen, T.; Voukantsis, D.; Niska, H.; Karatzas, K.; Kolehmainen, M. Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data. *Appl. Energy* **2010**, *87*, 3538–3545. [CrossRef]
11. Gouveia, J.P.; Seixas, J. Unraveling electricity consumption profiles in households through clusters: Combining smart meters and door-to-door surveys. *Energy Build.* **2016**, *116*, 666–676. [CrossRef]
12. Coke, G.; Tsao, M. Random effects mixture models for clustering electrical load series. *J. Time Ser. Anal.* **2010**, *31*, 451–464. [CrossRef]
13. Granell, R.; Axon, C.J.; Wallom, D.C.H. Impacts of Raw Data Temporal Resolution Using Selected Clustering Methods on Residential Electricity Load Profiles. *IEEE Trans. Power Syst.* **2015**, *30*, 3217–3224. [CrossRef]
14. Chicco, G.; Napoli, R.; Piglione, F. Comparisons Among Clustering Techniques for Electricity Customer Classification. *IEEE Trans. Power Syst.* **2006**, *21*, 933–940. [CrossRef]
15. Chicco, G.; Sumaili Akilimali, J. Renyi entropy-based classification of daily electrical load patterns. *IET Gener. Transm. Distrib.* **2010**, *4*, 736–745. [CrossRef]
16. Mcloughlin, F.; Duffy, A.; Conlon, M. A clustering approach to domestic electricity load profile characterisation using smart metering data. *Appl. Energy* **2015**, *141*, 190–199. [CrossRef]
17. Ndiaye, D.; Gabriel, K. Principal component analysis of the electricity consumption in residential dwellings. *Energy Build.* **2011**, *43*, 446–453. [CrossRef]
18. Kavousian, A.; Rajagopal, R.; Fischer, M. Determinants of residential electricity consumption: Using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants' behavior. *Energy* **2013**, *55*, 184–194. [CrossRef]
19. Tureczek, A. Clustering District Heat Exchange Stations Using Smart Meter Consumption Data. In Proceedings of the 3rd International Conference on Smart Meter Energy Systems and 4th Generation District Heating, Copenhagen, Denmark, 13 August 2017; p. 24.
20. Billard, L.; Douzal-Chouakria, A.; Samadi, S.Y. An Exploratory Analysis of Multiple Multivariate Time Series. In Proceedings of the 1st International Workshop Advanced Analytics Learning on Temporal Data AALTD 2015, Porto, Portugal, 11 September 2015; Volume 3, pp. 1–8.
21. Serrà, J.; Arcos, J.L. An empirical evaluation of similarity measures for time series classification. *Knowl.-Based Syst.* **2014**, *67*, 305–314. [CrossRef]

22. Carpaneto, E.; Chicco, G.; Napoli, R.; Scutariu, M. Electricity customer classification using frequency—Domain load pattern data. *Int. J. Electr. Power Energy Syst.* **2006**, *28*, 13–20. [CrossRef]
23. Chicco, G.; Ionel, O.M.; Porumb, R. Electrical load pattern grouping based on centroid model with ant colony clustering. *IEEE Trans. Power Syst.* **2013**, *28*, 1706–1715. [CrossRef]
24. Kang, J.; Lee, J. Electricity Customer Clustering Following Experts' Principle for Demand Response Applications. *Energies* **2015**, *8*, 12242–12265. [CrossRef]
25. Viegas, J.L.; Vieira, S.M.; Melício, R.; Mendes, V.M.F.; Sousa, J.M.C. Classification of new electricity customers based on surveys and smart metering data Commission for Energy Regulation. *Energy* **2016**, *107*, 804–817. [CrossRef]
26. Ramos, S.; Duarte, J.M.; Duarte, F.J.; Vale, Z. A data-mining-based methodology to support MV electricity customers' characterization. *Energy Build.* **2015**, *91*, 16–25. [CrossRef]
27. Liu, X.; Nielsen, P.S. Regression-based Online Anomaly Detection for Smart Grid Data. *arXiv*, **2016**, in press.
28. Lattin, J.; Carrol, J.D.; Green, P.E. *Analyzing Multivariate Data*, 1st ed.; Thomson Brooks/Cole: Duxbury, MA, USA, 2004; Volume 46.
29. Friedman, J.; Hastie, T. *The Elements of Statistical Learning*, 1st ed.; Springer: Berlin/Heidelberg, Germany, 2008.
30. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2012**, *12*, 2825–2830.
31. Al-otaibi, R.; Jin, N.; Wilcox, T.; Flach, P. Feature Construction and Calibration for Clustering Daily Load Curves from Smart-Meter Data. *IEEE Trans. Ind. Inform.* **2016**, *12*, 645–654. [CrossRef]
32. Perry, P.O. Cross-Validation for Unsupervised Learning. *arXiv*, **2009**, in press.
33. Madsen, H. *Time Series Analysis*; Chapman & Hall/CRC: Boca Raton, FL, USA, 2008.
34. Wasserman, L. *All of Statistics*; Springer: Berlin/Heidelberg, Germany, 2003; Volume C.
35. Barford, L.A.; Fazzio, R.S.; Smith, D.R. *An Introduction to Wavelets*; Technical Report HPL-92-124; Hewlett-Packard Labs: Bristol, UK, 1992; Volume 2, pp. 1–29.
36. Morchen, F. *Time Series Feature Extraction for Data Mining Using DWT and DFT*; 2003; pp. 1–31. Available online: http://www.mybytes.de/papers/moerchen03time.pdf (accessed on 22 February 2018).
37. Li, T.; Li, Q.; Zhu, S.; Ogihara, M. A survey on wavelet applications in data mining. *ACM SIGKDD Explor. Newsl.* **2002**, *4*, 49–68. [CrossRef]
38. Wasserman, L. *All of Nonparametric Statistics*; Springer: New York, NY, USA, 2006. [CrossRef]
39. Wasilewski, F. *PyWavelets*; 2006; Available online: https://pywavelets.readthedocs.io/en/latest/ (accessed on 22 February 2018).
40. Arthur, D.; Vassilvitskii, S. How slow is the k-means method? In Proceedings of the Twenty-Second Annual Symposium on Computational Geomeometry, Sedona, AZ, USA, 5–7 June 2006; p. 144.

Paper 3 - Clustering District Heat-Exchange Stations Using Smart-Meter Consumption Data

Clustering District Heat Exchange Stations Using Smart Meter Consumption Data

Alexander Martin Tureczek[a,*], Per Sieverts Nielsen[a], Henrik Madsen[b], Adam Brun[c]

[a]Systems Analysis, Management Engineering, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark
*atur@dtu.dk, Tel: +45-2346-0989, Produktionstorvet Building 426, room 125.
[a]Systems Analysis, Management Engineering, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark
pernn@dtu.dk
[b]Dynamical Systems, Compute, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark, hmad@dtu.dk
[c]Business Development, Affald Varme Aarhus, 8210 Aarhus V, Denmark, adbr@aarhus.dk

Abstract

Contrary to electricity smart meter data analysis, little research regarding district heat smart meter data has been published. Previous papers on smart meter data analytics have not investigated autocorrelation in smart meter data. This paper examines district heat smart meter data from the largest district heat supplier in Denmark and autocorrelation is identified in the data. The K-Means algorithm is not able to take autocorrelation into account when clustering. We propose different data transformation methods to enable K-Means to account for this autocorrelation information in the data by using wavelet transformation and autocorrelation features. Our results show that the K-Means yield acceptable clustering results for district heat data when clustering normalized data, inclusion of autocorrelation improves the clustering. The clusters on normalized data are similar to the wavelet transformed clusters, where the autocorrelation has been accounted for. The clustering achieved with the autocorrelation transformation yields finer clusters through accounting for autocorrelation. We are not able to statistically show a difference between the transformations. All transformations result in shadowing clusters, but the autocorrelation transformation generates fewer shadow clusters and reduce the number of dimensions from 744 to 24, resulting in a dramatic reduction in K-Means runtime.

Highlights

1. Existence of autocorrelation in the smart meter data is shown.
2. Preprocessing of data enables K-Means to account for autocorrelation.
3. Implementation of cross-validation for unsupervised learning.
4. Autocorrelation feature from input data enable finer clustering from K-Means.
5. Wavelets reduce data with ensuing faster runtime without notable change to clusters.

Keywords

Clustering; Feature Extraction; Autocorrelation; Wavelet Analysis; Smart Meter Data; Load Pattern;

1. Introduction

In traditional electricity production, fossil fuels and organic waste is burned to generate electricity, the incineration process generates heat as a bi-product. In Denmark, this heat is increasingly utilized as a source for domestic and industrial heating, through district heating systems. Combined heat and power plants (CHP) help cover the demand for electricity and heat of larger populated areas, while smaller heat production plants supply rural districts Overall District Heating supplies heating and hot water to approximately 64% [1] of all Danish households.

Since the energy crisis in 1973 district heating has received much attention in Denmark as an efficient and environmentally friendly means to ensure heating. Current research in district heating systems is culminating with 4th generation, where the dominant research focus has been directed towards more efficient technical solutions such as better pipes, flow control, drag minimizing additives and lower temperature in the system. Much less effort has been directed towards understanding consumption as part of the system and driver of the demand.

Recent advent of smart meters, for automatic metering of consumption, has enabled consumption recordings at an unprecedented detail, moving from biannual manual readings to automatic hourly readings. Many Danish district heat utilities have embraced this technology and installed smart meters at heat exchange stations, where the transmission and distribution grid transfer energy to individual household areas. Installation of meters at these levels enable the district heat utilities to supervise the systems in an intelligent way, but also to increase their knowledge of their consumers for better understanding the demand.

The past decade has witnessed a strong research and technology focus on electricity smart grids where the introduction of smart meters at household level has enabled detailed recording of consumption. Recording of consumption at household level by the minute can help optimize the electricity grid and identify flexibility in the entire system, by way of increased consumer knowledge. Introduction of modern metering equipment for district heating, will in theory, make it possible to do the same consumption pattern analysis, which successfully has been applied to electricity smart meter data in the last decade. This paper will investigate the feasibility of the learnings and methods from smart meter electricity consumption clustering in a district heat setting and assess if these learnings are readily applicable to district heating consumption data.

Heat smart meter data analysis is not only relevant for understanding heat consumption patterns in district heating networks from an academic point of view, but also allow the district heating companies to understand their consumer better and potentially optimize their heat delivery service. We suggest that heat smart meter profiles can be used at a generic level to model heat consumption patterns in areas outside of the district heating district, where heating is based on individual heating sources. In this way the heat smart meter data can be used to model local air pollution levels and potentially forecast air pollution levels in these areas using weather forecast data such as temperature and wind forecasts.

This paper makes the following contributions: First, this paper presents analysis of smart meter consumption data at heat exchange station level, including imputation of outliers. Second, the paper confirms the existence

of autocorrelation in smart meter data, information which K-Means ignores in the clustering. Third, the paper introduces methods to enable K-Means to account for autocorrelation information in data, by careful transformation of K-Means input data. This can easily be extended to account for other intrinsic data structures. Fourth, the paper extends the notion of cross-validation into unsupervised learning by creating pseudo response variables from cluster validation indices.

The rest of this paper is divided into 5 sections; current relevant literature is listed in section 2, while section 3 describes the data and the preprocessing performed, section 4 discusses the methods applied for clustering. In section 5 the results from clustering using K-means, with and without transformed data are shown. Section 6 discusses the results and their implications while section 7 concludes upon the findings. Finally section 8 includes an executive summary.

2. Literature Review

A literature study [2] on smart meter data clustering, which evaluated more than 2000 papers concerning application of energy smart meter data for consumption clustering, identified no papers applying district heating smart meter data for household consumption clustering. Electricity consumption smart metering data have demonstrated significant differences in consumption patterns [3]. This discrepancy can be attributed to differences in lifestyles illustrating the need for better understanding of consumer behavior and consumption patterns, to facilitate more efficient use of resources.

The review [2] also supplies a list of prevalent methods for consumer clustering applying smart meter data. K-means is frequently, and thoroughly studied [3–5], and repeatedly compared to more advanced methods such as follow-the-leader [6,7] and hierarchical clustering [8–10]. Experiments with data transformation for preprocessing input data prior to K-Means clustering is conducted in [11]. Evaluation of the resulting clusters is estimated by applying various validation indices [12–14].

There was no clear indication of time interval to evaluate ranging from days [15] to a year [16], and few papers amended external data such as weather data [8], survey information [17,18] and occasionally with energy audits [19] to evaluate consumers and behavior [20]. Only one recent paper which applied smart meter data on district heating was identified; [21] applies hourly meter data from district heat substations to evaluate heat load patterns on predefined customer classes.

With only few identified studies relating to district heating smart meter data, we will draw much inspiration from research done in electricity smart meter analysis, where K-Means is extensively applied for clustering. The simplicity of the K-Means algorithm has some caveat to which we will discuss solutions. We hypothesize the existence of autocorrelation in the smart meter data. Existence of autocorrelation has consequences for K-Means ability to cluster. Confirmation of the existence of autocorrelation will require circumvention by carefully preprocessing the input data to the K-Means algorithm leaving the K-Means algorithm unaltered.

3. Data Summary and Preprocessing

This section describes the data analyzed in this paper, we do so by applying a data description table proposed by [2], presenting important data information in a table view. Furthermore this section will describe the data preprocessing completed while preparing the data for analytics.

The data applied in this study is kindly supplied by AffaldVarme Aarhus (AVA), the largest district heat supplier in Denmark, covering the municipality of Aarhus including suburban and rural areas. Additionally, AVA also supplies a handful of smaller municipalities such as Skanderborg, Odder, Hørning and Hornslet through their transmission grid. The raw data initially comprises hourly readings from the 53 heat exchange stations, which we hereafter refer to as "HX stations". HX stations link the pressurized high temperature water in the transmission grid to the lower temperature distribution grid. The data recordings encompass the time period January 1st 2017 till January 31st 2017. Three HX stations have been removed from the data set because of incomplete readings. The data description in Table 1 gives an overview of the data set, with the initial and final sample sizes applied in this paper. The structure was proposed in [2] as a standardized means to report minimum data set information to the reader.

Data Description	Value
Type	Smart meter readings from district heat exchange stations, exchanging heat from transmission to distribution grid. Supplying smaller geographical areas of residential and industrial consumers with heat.
Country	Denmark
Region	Municipality of Aarhus
Supplier	AffaldVarme Aarhus (AVA)
Initial Data Size	53 district heat smart meters with 744 observations each.
Exclusion of Data	Meter **#130** was removed as it is a large company Heat exchange station servicing only 1 customer.
Missing Values	Meters: **#118A**, **#136J**, **#147** discarded du to missing data for entire January. Meters: **#111C**, **#119**, **#133**, **#134**, **#135**, **#136**, **#148** and **#151** erroneous readings. These errors were imputed as described in section 3, and the meters kept in the data set.
Final Data Size	49 district heat smart meters with complete data
Recording Frequency	Hourly
Start	01/01/2017
End	31/01/2017
Length	744 recordings per meter. Hourly recording for entire January.
Referral	Data never before referenced.

Table 1 - Data description as proposed in [2]. The table produces an overview of the data utilized in this study; how many smart meter readings are included in the final data set and which smart meters were disregarded for what reason.

The data set from AVA contains, as shown in Table 1, meter readings from January 1st, 2017 until January 31st, 2017, yielding 744 recordings per HX station. January is one of the coldest months in Denmark and thus a period where the HX stations are very active the entire day and month as most customers demand heat.

The HX meter readings utilized in this study are complete for most readings. Three HX stations (118A, 136J, 147) have missing values for the entire January, and are removed from the dataset; the three HX stations are one mobile HX station, and 2 HX stations supplying an external utility company. Finally, HX station 130, was removed as it is a station supplying only one company and hence not representative as an HX station.

Meters 111C, 119, 133, 134, 135, 136, 148 and 151 all experienced faulty readings, seen as spikes or outtakes in Figure 1 (left). The erroneous readings are easily identified visually, but also via data, as they all have zero-consumption readings. Furthermore a few of the meters exhibit sudden very high consumption readings. There are several scenarios which can explain the erratic readings; service stops, cold water pockets, or pipe failure. AVA registered none of those in the given period hence we treat the outliers as meter misreading. Rectification of the outliers is done by imputing the series mean value onto each outlier. This is a simple technique and proved very successful with the current data as can be visually inspected in Figure 1 (right) where the error mean corrected series are shown. All spikes and breaks in the meter readings have been evened out without inducing noticeable artifacts.

Figure 1 - (left) Original meter readings from HX stations with clear recording errors. Affected HX stations 111C, 119, 133, 134, 135, 136, 148 and 151. (Right) Correction of errors by imputing series mean into the faulty reading.

No other concerns were observed in the data; hence the error mean correction is the only data manipulation applied to the data before analysis, resulting in a final sample size of 49 meters each with 744 readings encompassing January 2017. A recent study analyzing district heat end user consumption investigates household smart meters from the same region whereas this study analyzes district heat exchange stations servicing larger areas [22].

4. Methodology

This section will outline the methods applied to the data. Starting with a description of the clustering technique K-Means in section 4.1, section 4.2 will introduce data scaling, which is needed for K-Means. Section 4.3 describes the selection of the number of clusters. Cross-validation is introduced in section 4.4. Finally section 4.5 and 4.6 introduces transformations of the input data by way of autocorrelation functions and wavelets.

4.1 K-means Clustering

Clustering HX stations into smaller homogenous subsets for better encapsulating consumption structures is an unsupervised problem. There is no prior knowledge of the true underlying clusters and hence no known structure to model against. In comparison, supervised classification utilizes a response variable, usually denoted Y, and a link function to model the underlying structure. From previous studies of electricity smart meter clustering [2] a considerable amount of clustering is performed by applying K-means or derivatives of the K-means algorithm. Consequently K-means will act as a benchmark for the analysis performed in this paper. K-means has some properties which make it a popular choice for unsupervised clustering, but there are some caveats which must be addressed.

K-means is readily implemented in most modern analysis software from proprietary to open source. In absence the algorithm is simple to implement. The algorithm is very robust and usually able to cluster data satisfactory. There exist K-means derivatives which are optimized for handling outliers or allowing fuzzy class-labels. The simplicity of the algorithm is its raison d'etre, in comparison to more advanced algorithms which do not necessarily perform significantly better. This further propels the popularity of general purpose algorithms like the K-means algorithm for unsupervised clustering.

There are some significant caveats concerning the K-means method, most significant is its greedy design philosophy [23] combined with the random initialization of the algorithm. The algorithm is prone to identify locally optimal solutions, which can lead to results that are not globally supported. To alleviate this problem the algorithm is usually run several times with random initialization and then selecting the best solution among the runs. This paper utilizes the K-Means implementation found in the scikit-learn framework for python [24] which by default selects the best clustering of 10 repetitions. Unless the initial random seed is fixed, the algorithm will return new clusters each run. Throughout this paper the random seed is set to 12345 to ensure reproducibility.

The algorithm does not account for any inherent structure in the data, for instance autocorrelation often found in time series. Either the algorithm is improved to handle more data structures or the input data can be preprocessed in such a fashion that the algorithm can cope with the inherent information in specific data structures. In this paper, we will pursue preprocessing of data to improve clustering performance.

The algorithm is initialized with k number of clusters; the clusters can be random or predefined. Repeatedly each observation is assigned the cluster which is closest to the observation, after which the cluster centers are recalculated including the newly assigned observation. This process is continued until no change occurs in

cluster assignments and the algorithm has converged to at least a local optimum. For an intuitive and thorough discussion of the mechanics of the algorithm see [25,26].

4.2 Data Scaling and translation

The K-means algorithm compares data by difference, grouping HX stations with large energy throughput together regardless of their consumption pattern. To avoid classifying the amount of energy consumed rather than the pattern by which it was consumed, we need to remove scale from the data. Similarly [11] applied scaling as preprocessing of smart meter data. In this paper we apply four different scaling and two transformations. The transformations are intended to amplify the difference between the groups, which makes it easier to recognize homogeneous clusters.

Scale	Mathematical Description	Intuition
Normalization	$\frac{x - x_{min}}{x_{max} - x_{min}}$	Normalization puts all observations on a 0-1 scale compared to the largest reading. Dimensionless.
Standardization	$\frac{x - x_{mean}}{\sigma}$	Standardization scales all observations compared to the variance of the time series. Dimensionless.
Mean-Center	$x - x_{mean}$	Mean-centering removes the mean from the meter reading. It is equal to shifting on the y-axis.
Mean-Divide	$\frac{x}{x_{mean}}$	Scales observations relative to the series mean. Does not constrain the y-axis to the interval [0, 1]. Dimensionless.

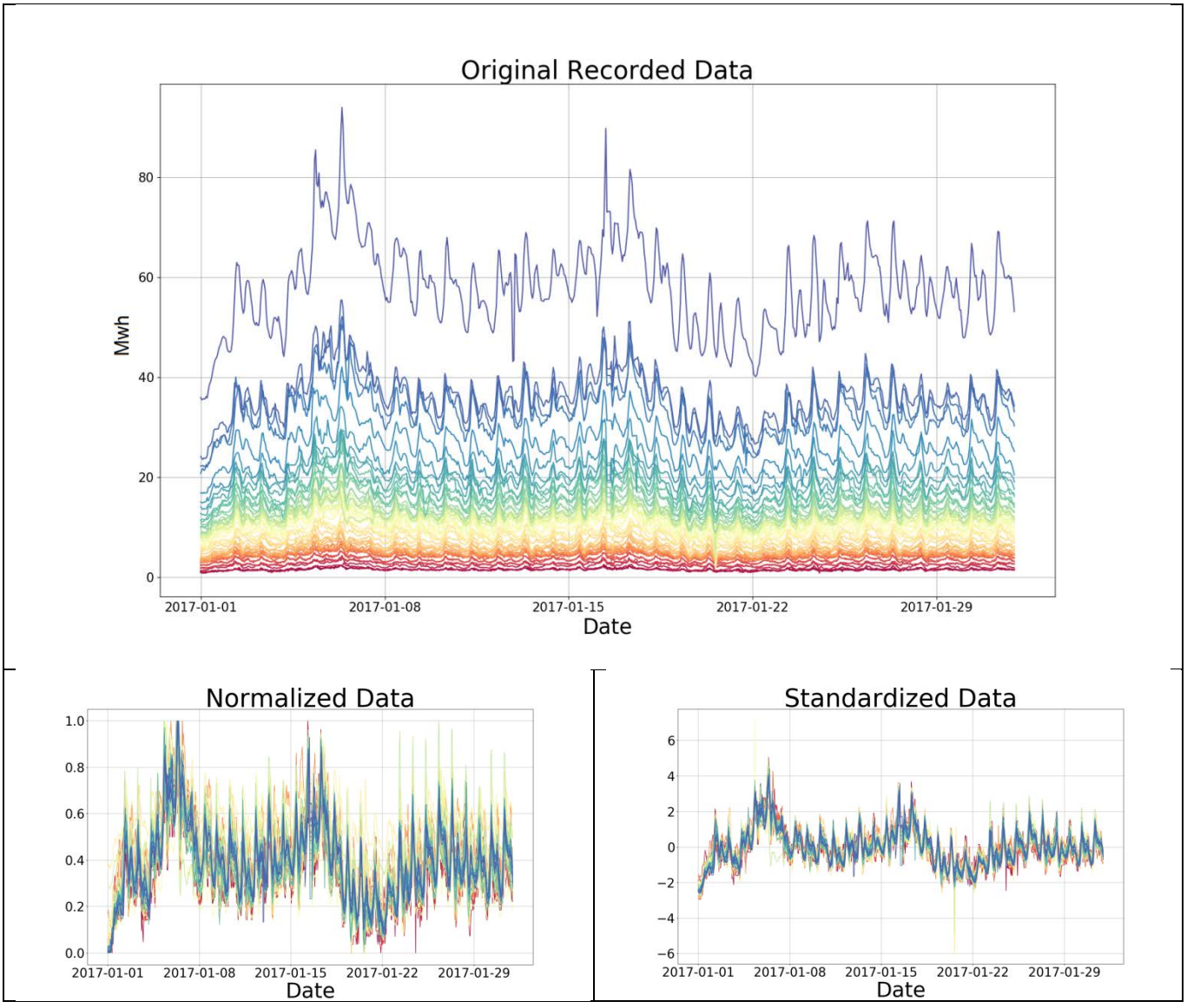
Table 2 - Scaling methods applied in this paper, to remove consumption volume impact from HX stations. Scaling can help reveal true consumption patterns and disregard volume throughput differences of individual HX stations.

The 4 different scaling techniques listed in Table 2 all remove the differences in consumption volume while retaining the consumption patterns of each HX station. *Normalization* scales readings from each HX station to the interval [0:1]. *Standardization* scales the consumption patterns with respect to standard deviation of the HX station with mean equal 0 and unit variance. *Mean-Centering* scales by removing the mean from the HX station, shifting the mean to 0. Mean-divide is comparable to normalization, dividing by series mean. It does not constrain the y-axis to the interval [0, 1].

The scaling methods remove volume differences, through different strategies, essentially retaining only the pattern in the data. Scaling can be essential in clustering techniques where clustering is performed on a distance metric. As mentioned K-means is prone to cluster HX stations of equal consumption volume regardless of difference in consumption pattern. The data are scaled to ensure HX stations are clustered by pattern and not consumption volume. Figure 2 (top) shows the original data and the implications of the different scaling applied (bottom), which illustrate that removing consumption volume difference of the HX stations reveals similar consumption structure across the stations.

213

214


Original Recorded Data
Mwh
0
20
40
60
80
2017-01-01
2017-01-08
2017-01-15
2017-01-22
2017-01-29
Date
Normalized Data
0.0
0.2
0.4
0.6
0.8
1.0
2017-01-01
2017-01-08
2017-01-15
2017-01-22
2017-01-29
Date
Standardized Data
−6
−4
−2
0
2
4
6
2017-01-01
2017-01-08
2017-01-15
2017-01-22
2017-01-29
Date

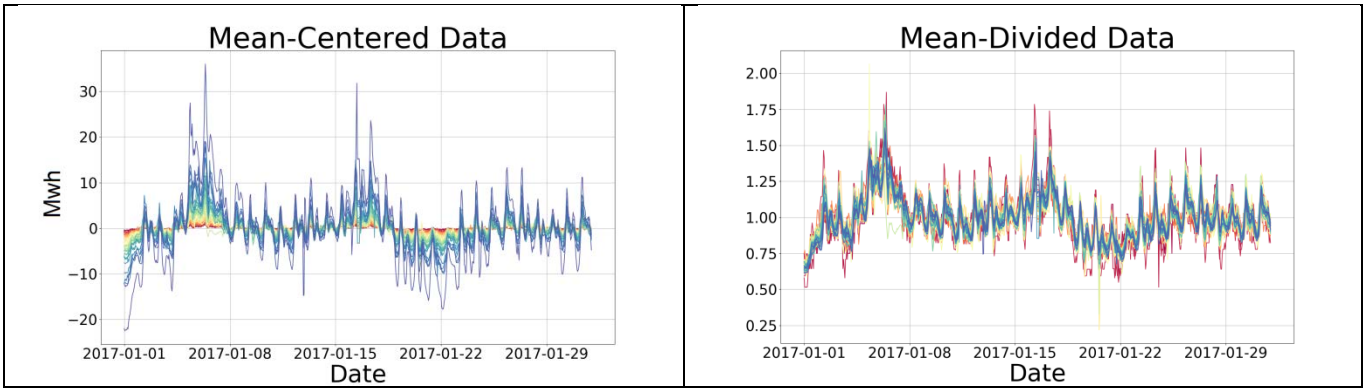


Figure 2 - Top: the original data plotted, clearly showing a difference in consumption volume rather than different consumption patterns. Mid-Left: normalizing data reveals the similar consumption pattern for each HX station. Likewise does standardization, mean-centering and mean-dividing of the data. The scaling reveal similar consumption pattern across all 49 HX stations with different consumption volume. The coloring shows the small differences in the different HX stations consumption after scaling.

4.3 Selecting the Number of Clusters

K-Means does not give a solution to selecting the optimal number of clusters. It purely classifies according to the initial user-selected number of classes and the random seed. In order to select the optimal number of clusters, we need a metric for evaluating different clustering solutions, without knowing the true underlying clusters. A multitude of different measures, for assessing clustering performance have been developed to help identify the optimal number of clusters. These metrics are concerned with quantifying inter and intra variability of the resulting clusters. This paper will employ 4 different cluster validation indices: MIA, Cluster Dispersion Indicator, Davies-Boudin Index and the Silhouette index. All indices applied are described in Table 3 which is an adaptation of a table presented in [2].

Index	Mathematical description	Properties
MIA	$$\boldsymbol{MIA} = \sqrt{\frac{\mathbf{1}}{\boldsymbol{K}}\sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{d^2(C_k)}},$$	Average distance within class-member to class centroid, summarized across all classes. k is number of clusters; $d^2(C_k)$ is the squared average distance within cluster k. High MIA indicates large distances within the classes. E.g., large dispersion.
Cluster Dispersion Index (CDI)	$$\boldsymbol{CDI} = \frac{\mathbf{1}}{\boldsymbol{d(C)}}\sqrt{\frac{\mathbf{1}}{\boldsymbol{K}}\sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{d^2(C_k)}}$$	CDI prefers Long inter-cluster distance and short intra-cluster distance [14]. Small values indicate good clustering. $d^2(C_k)$ is the squared average distance within cluster k. While $d(C)$ is average cluster distance in data [14].
Davies-Boudin Index (DBI)	$$\boldsymbol{DBI} = \frac{\mathbf{1}}{\boldsymbol{K}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{K}} \boldsymbol{max}_{\boldsymbol{j\neq i}} \frac{\boldsymbol{diam(C_i)} + \boldsymbol{diam(C_j)}}{\boldsymbol{d(C_i, C_j)}}$$	$diam(C_k)$ is the average diameter of a cluster. And $d(C_i, C_j)$ is the distance between cluster centers. K is the number of clusters. DBI relates the mean distance of each class with the distance to the closest class [27]. Smaller values of DBI implies that K-means clustering algorithm separates the data set properly [16]
Silhouette Index	$$\boldsymbol{Silhouette} = \frac{\boldsymbol{c'(x)} - \boldsymbol{c(x)}}{\boldsymbol{max}\{\boldsymbol{c(x)}, \boldsymbol{c'(x)}\}}$$ $$\boldsymbol{c'(x)} = \boldsymbol{\min_{y\in C'}} \boldsymbol{d(x,y)}$$	$c(x)$ is the average distance between vector x and all other vectors of the cluster c to which x belongs. $c'(x)$ is the minimum distance between vector x and all other vectors in cluster $\forall\ C' \neq C$ [11]. SI is between [–1, 1] higher is better. Negative is miss-

		clustering

Table 3 - Overview table of Cluster validation indices applied in this paper, with their mathematical description and intuitive properties. The table is an adaptation from a table in [2].

Like residuals and r^2 in regression analysis, these indices are developed so as to minimize the dispersion within and maximize the distance between clusters, helping to select the optimal number of clusters. Plotting the progression of the indices as function of clusters allows for visual inspection, where abrupt changes in their decline or fluctuating pattern can help select the number of clusters within a given data set. This method of estimating the number of clusters in unsupervised clusters has been studied by [28].

4.4 Cross-Validation

Cross-validation is a concept developed for supervised learning [29] as a bias-variance trade-off for reducing misclassification by splitting the data into a test and training set [30]. In the test set the true cluster label is known, which enables comparison of the model clustering vs the true clusters.

We will apply cross-validation, not directly to the unknown clusters but to the cluster-validation indices thus regarding the indices as pseudo cluster labels. We apply leave-one-out cross-validation [31], calculating the indices for all 49 HX stations with the omission of one. This is done repeatedly until indices have been calculated for each combination of 48 HX stations. We report the maximum, minimum and average index value, of all four validation indices.

4.5 Autocorrelation Feature Extraction

The HX station meter readings are ordered in time, with equidistant intervals, each HX stations data output can be regardeds as a time series. An extensive literature exists on Time Series Analysis which we will not cover in this paper. For a thorough discussion refer to [32], for a survey on time series clustering with discussion on difference in type of methods see [33]. An important tool from classical Time Series Analysis is the autocorrelation function (ACF) which is defined as:

$$ACF(\tau) = \frac{E([X_t - \mu][X_{t+\tau} - \mu])}{\sigma^2} \quad (1)$$

where t and τ are integer time steps, μ is the series mean, and σ^2 it the series variance. Intuitively the data generating process is correlated with itself, quantifying how much previous observations influence the current observation. The n lags in an ACF plot indicate the influence of each of the previous n observations and they are essential in identifying the type of time series process which generated the readings. The ACF plot visualizes inherent information such as dampening, oscillation, and recurrences throughout the data. The data are expected to contain autocorrelation as the output from the HX stations is expected to exhibit a daily cycle. To the best of our knowledge this has only briefly been studied in smart meter electricity analytics by [20] and has not been utilized as input for K-Means clustering in a smart meter setting. In this paper we will not perform a rigid time series analysis of each HX station, nor will we develop ARIMA models for the individual HX Stations.

As we do not develop ARIMA models we do not need to ensure stationarity of the time series. We will merely evaluate differences in ACF and apply the coefficients as input for the K-Means algorithm to cluster.

K-means is by default not leveraging the expected auto-correlation information in the HX station readings as mentioned in section 4.1. We preprocess the input data by calculating the ACF coefficients of each HX station, and input the resulting ACF coefficient matrix into the K-Means algorithm thereby enabling the K-Means to account for the autocorrelation in the data. K-Means is enabled to account for auto-correlation by classifying the ACF structure rather than the observed consumption data from each station. As the ACF is employed to characterize the underlying model and parameters, clustering the ACF coefficients can be regarded as clustering with respect to the underlying process. Under certain conditions a rigid time series analysis and model can be developed describing the individual clusters.

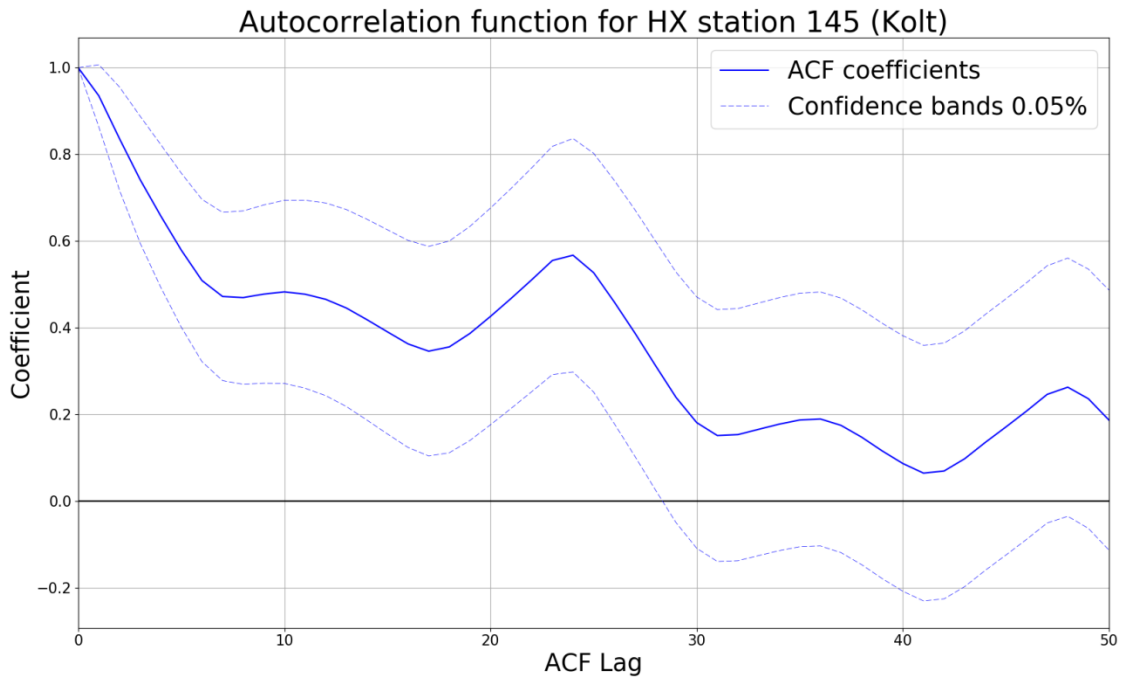


Figure 3 - Autocorrelation plot with 50 lags of HX station 145 the town of Kolt. The 0.95% confidence bands shows significant lag coefficients until lag 28. A clear seasonality is also seen at lag 24 indicating a daily recurrent pattern.

Figure 3 shows a 50-lag autocorrelation plot including 95% confidence intervals of the HX station; 145 Kolt. The Confidence intervals shows the first 28 lag coefficients to be significantly different from 0. The Figure confirms the existence of autocorrelation in the data. Moreover the Figure shows a seasonality component every 24 lag, indicating a recurrent daily cycle. The ACF is invariant to the scaling and translation described in section 4.2 meaning volume differences are irrelevant when ACF preprocessing the data.

4.6 Wavelet Feature Extraction

The wavelet transformation is a basis transformation using wavelet basis functions. Wavelets are able to represent smooth and locally non-smooth functions. Wavelets have time and frequency localization, effectively linking time and frequency in contrast to the Fourier transformation which only allows frequency localization [31]. Wavelet are especially well suited for analyzing high frequency data because of their ability to capture global smoothness and local spikes in the signal [30], while filtering out high frequency noise [34]. The application of wavelets for time series feature extraction in this study has been inspired by [35]. In the process of filtering high frequency data, wavelets perform efficient data compression, by removing non-significant coefficients. Often this process removes a considerable number of coefficients. The decomposition of the signal into wavelet coefficients are not easily human interpretable, but are readily applicable as input for the K-Means algorithm. The wavelet coefficients are uncorrelated [36].

A wavelet defines the orthonormal basis function, and is scaled and shifted to fit the signal. Choosing a suitable wavelet can be difficult, as the scaled basis wavelet must be able to encapsulate the structure of the signal. For the selection of the wavelet applied for analysis of the HX data refer to section 5.3. As wavelet definitions are often very complex, we here present a brief overview of the theory of wavelets using the Haar wavelet. The Haar wavelet is defined by [37]:

$$\psi(x) = \begin{cases} 1, & \text{if } 0 \leq x < \frac{1}{2} \\ -1, & \text{if } \frac{1}{2} \leq x < 1 \\ 0 & \text{Otherwise,} \end{cases} \tag{2}$$

and is shown in Figure 4. The basis function is scaled and shifted by the scaling function defined as:

$$\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k) \tag{3}$$

where j and k are integers. This function has the same shape as ψ but is scaled and shifted [30]. The wavelet is repeatedly applied at different scales to locally fit any fluctuating and smooth regions.

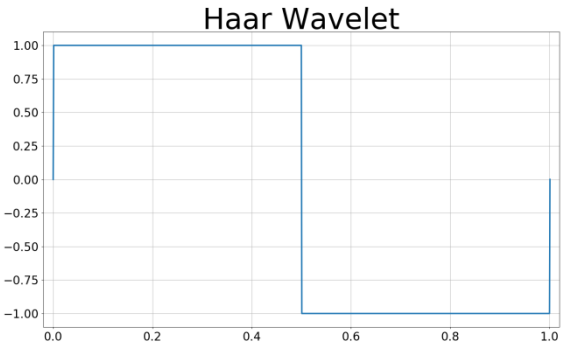


Figure 4 - Plot of the Haar wavelet in its area of definition.

Removing non-significant coefficients is called thresholding. We apply universal thresholding to the coefficients $D_{j,k}$ which are kept as wavelets parameters $\hat{\beta}_{j,k}$. The coefficient is evaluated for statistical significance by:

$$\hat{\beta}_{j,k} = \begin{cases} D_{j,k} & \text{if } |D_{j,k}| > \hat{\sigma}\sqrt{\frac{2*\log(n)}{n}} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

Wavelets other than the Haar wavelet are usually difficult to describe in closed form, but numeric tools for computation are available[30]. There exists a fast algorithm for wavelet coefficient calculation which processes in linear time [38], which in algorithmic analysis o-notation is equivalent to $O(n)$ [23]. The python wavelet package PyWt [39] was utilized for the wavelet analysis presented in this paper. Figure 5 (Left) shows a Haar wavelet transformation fit to the HX station; 145 Kolt. The original time series has been transformed via the wavelet, retaining only significant coefficients. The wavelet transformation has compressed the time series to 161 coefficients compared to the original data of 744 observations. The figure shows how the wavelets are able to keep the important structure in the data even at a 1:5 compression. The structure of the Haar wavelet results in its inability to truly follow the data, producing noise reduced representation of the original data. The corresponding Haar wavelet pyramid plot of HX station 145 Kolt is seen in Figure 5 (Right).

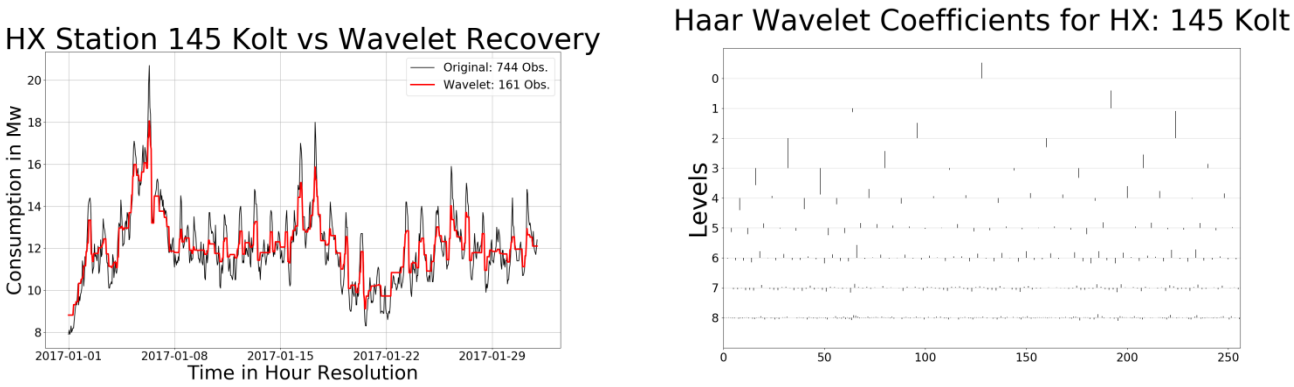


Figure 5 – (Left) Illustration of Haar wavelet approximation to original data series. The wavelet has a compression factor close to 1:5 and still recovers large parts of the structure in the original time series. Only 161 wavelet coefficients were significant different from zero, compared to the entire data series of 744 observations. (Right) Haar wavelet pyramid plot after applying Haar wavelet to the HX station 145 Kolt.

The visual analysis of wavelets is done using the pyramid plot. The algorithm compares neighboring observations pairwise e.g. in the small time series x = [1, 2, 3, 4] observation (1, 2) are compared, and (3, 4) are compared. This comparison yields the first level in the pyramid algorithm. Next pair (1, 2) and pair (3, 4) are compared, producing the second level, and so forth until there are no more pairs to compare. This can be plotted as a pyramid-plot with the lines indicating size of the difference from the comparison done at each level. Universal thresholding is then applied on the lowest level to remove noise. For clustering we generate new data series from the coefficients and cluster the HX stations by their wavelet coefficients.

5. Results

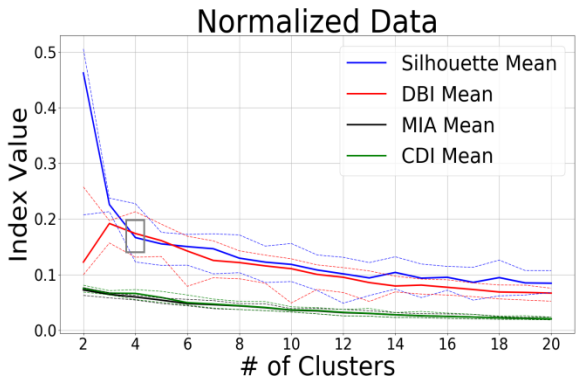
This section will present the results from K-Means clustering performed on scaled data in section 5.1, while section 5.2 compares clustering performance on autocorrelation transformed data, and subsequently in section 5.3 the results from the wavelet transformation is presented. Section 5.4 presents a comparison of the different clustering achieved

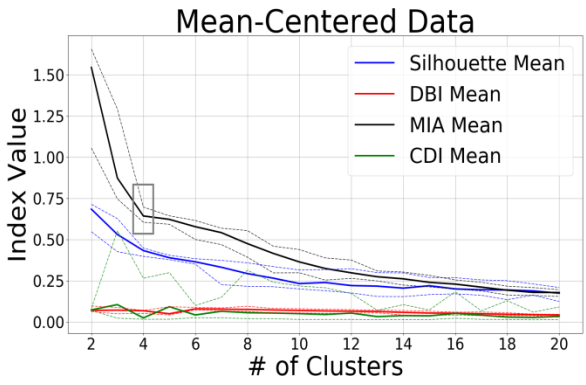
5.1 Cluster Performance: Normalized Data

We see how well we can classify HX stations with K-Means and whether careful transformation of the original data can improve the performance by applying scaling and ACF transformation to the input data. K-Means clustering done on scaled data constitutes our benchmark for comparing the input transformation influence on K-means clustering performance.

To the best of the authors' knowledge HX stations have never before been classified by applying smart meter recordings. Our hypothesis is that the data will behave similarly to electricity smart meter data, we therefore progress with the analysis of HX stations in identical manner as observed in [2] regarding electricity smart meter data.

We plot the cluster validation index to select the optimum number of clusters for the scaled data and study how they develop as the number of clusters increases from 2 to 20 clusters. All 4 validation indices; Silhouette, MIA, CDI and DBI are calculated for each of the scaling, Normalized, Standardized, Mean-Centered, and Mean-Divided. Figure 6 shows how the mean of 4 indices develops as more clusters are introduced. The dark grey box overlay indicates a potential optimum number of clusters.





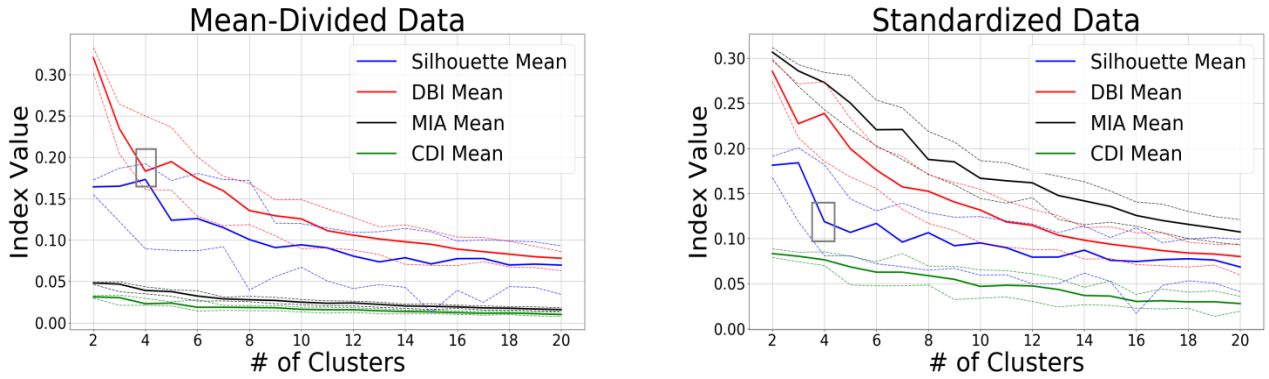


Figure 6 – Cluster number selection. Random Seed: 12345. Pseudo Cross-validation is performed on each of the validation indices. The maximum and minimum recorded value for each index and cluster number is shown with dashed lines, while the bold line is the average index value. The dark grey box indicates the optimum number of clusters for each scaling.

The silhouette index performs especially well for the normalized data, with an "elbow" break at four clusters indicating marginal performance increase by including more clusters. For the mean-centered data the MIA index shows an "elbow" break around four clusters. For the Mean-divided data the DBI indicate an "elbow" break occurs at four clusters, while the silhouette index is close to the DBI index. Finally the Standardized data shows a steady almost linear decline of the indices indicating no apparent cluster size cutoff. The CDI index is not indicating any number of clusters for any scaling following a linear decline in each case. Three different indices indicate four as the optimum number of clusters, which will be our choice of clusters.

It is not apparent from Figure 6 which of the scalings to apply. Only standardization falls short of giving an indication of how many clusters to include. We select normalization with four clusters as it yields the most balanced clusters where the smallest cluster has four members. In comparison the smallest cluster for mean-divide and mean-centered only has one member, Table 4 shows the cluster size after scaling.

Transform	Cluster size @ 4 clusters
Normalization	(18, 4, 12, 15)
Mean Divided	(5, 29, 14, 1)
Mean Centered	(21, 4, 23, 1)
Standardization	(11, 12, 22, 4)

Table 4- Resulting cluster sizes with 4 clusters and transformations; Normalized, Mean-Centered, Mean-Divided and standardization. Normalization yields the most balanced clsuters.

The resulting clustering; four clusters and normalized data, is shown in Figure 7 (left) where the 4 cluster means are plottet. The means exhibit the same general structure with a slight offset and it is possible to destinguish individual cluster means. Superimposing the individual cluster members onto the graph Figure 7 (right) visualizes the overlap between the classes. The clusters are seperated but not to such an extend that they are easily separable due to large within cluster variation resulting in overlapping clusters.

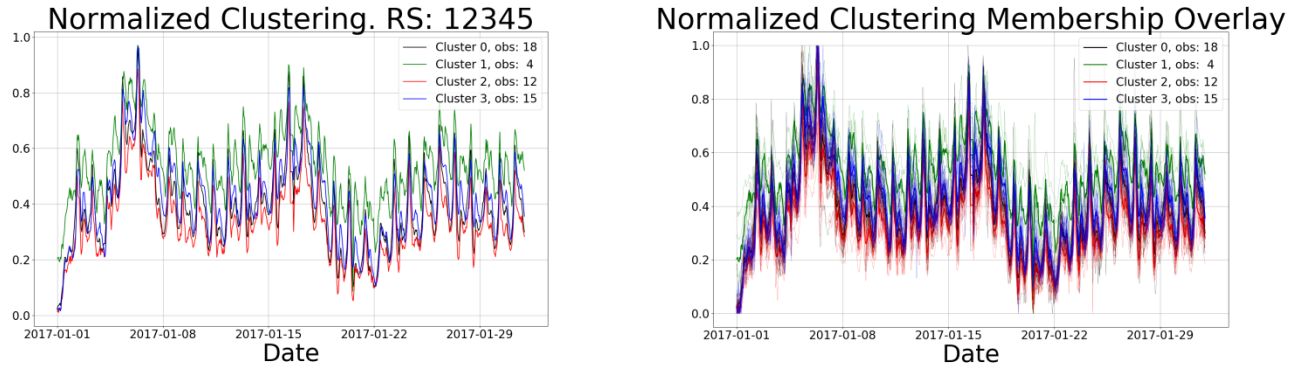


Figure 7 - (Left) 4 class means, after normalization. (Right) overlay of all class member onto the class means, cluster blue, red and black are very close while the green cluster is distingushable form the other 3 clusters.

This effectively renders the clustering an academic exercise that shows K-means can be applied to successfully identify clusters, but their true separability and practical applicability is questionable. The overlap of the clusters was expected due to the low index values seen in Figure 6, especially the CDI indicated clusters could be overlapping. The outcome does not change by changing the meter window from one month to weekly or daily basis which also results in overlapping classes.

The resulting clusters are closely located, which makes the individual clusters difficult to apply in a non-academic setting. We have to be more aware of the features of the input data to the K-Means algorithm to remedy the situation. Clever data transformation is needed before clustering to circumvent the weaknesses of the K-Means. One of these methods could be the application of the autocorrelation structure as input features, enabling the K-Means to treat auto-correlated data when clustering. Figure 8 shows how the clusters are distributed across the consumption volume and not grouped with equal consumption clusters.

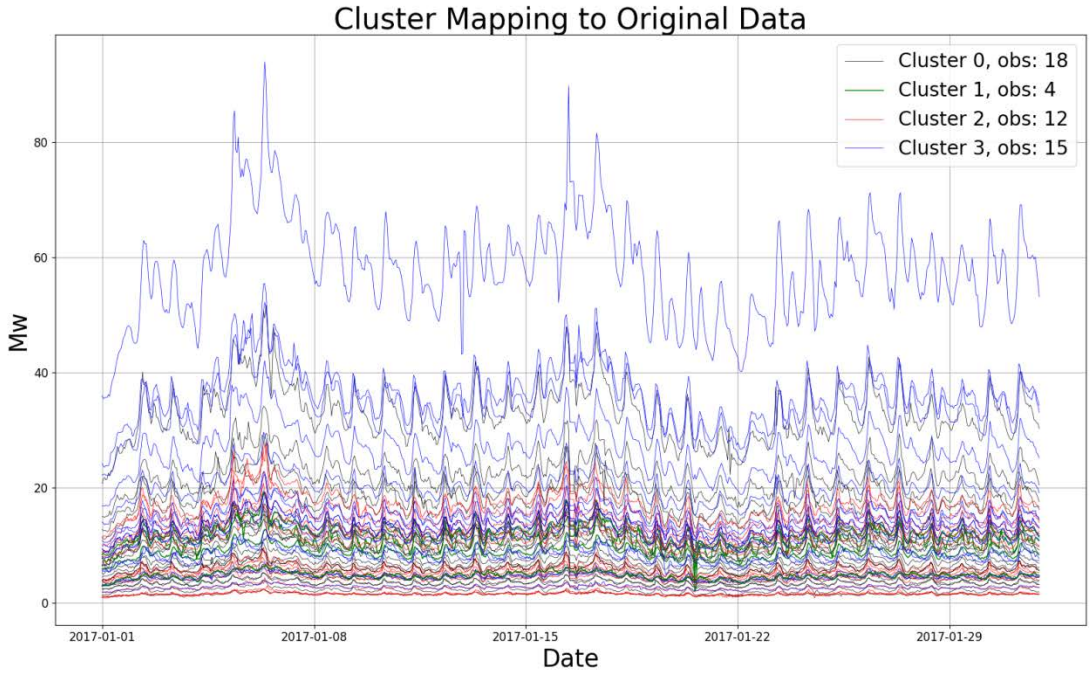


Figure 8 - Mapping of class members to original data. It is clearly seen that the clusters are not biased by consumption volume.

Clustering by applying scaled data ensures the clusters are not biased by differences in consumption volume of individual HX stations. Scaling alone does not utilize latent information inherent in the data. Contrarily, the ACF is invariant to the volume and focuses only on the underlying consumption structure of each HX station which makes scaling of the data irrelevant. Additionally the ACF also cater important information for Box-Jenkins classical analysis of time series and thus points the way for deeper knowledge of the underlying process that generated the consumption in each cluster through further analysis.

5.2 Cluster Performance: Autocorrelation Feature Extraction

To enable K-Means to account for autocorrelation in the data, we calculate the autocorrelation coefficients of all HX stations with 24 hours lag, and apply this as input to K-Means. Only significant autocorrelation coefficients are included in the input data set to ensure statistical stability. 24 hour lag has been chosen as the basis for the general structure encompassing one daily cycle. Figure 3 shows that coefficients above 24 lags become non-significant, with the recurrent pattern seen in the figure continuing for multiple lag beyond 24 lags. This pattern of significance was observed for all the HX stations.

The exact same method as described for clustering of scaled data has been applied to develop the 4 cluster validation indices as a function of number of clusters from 2 to 20 illustrated in Figure 9 (Left). There is a clear change in progression near 7 clusters. Figure 9 (right) shows the different cluster means for 6 clusters, the 7th cluster has only two observations and hence is not plotted. The resulting 6 clusters show visible differences and are more separable than the scaled clusters seen in Figure 7 (left). The separability of clusters is especially

noticeable in the region lag 8 to lag 17, while at the end points the clusters exhibit similar structure, Figure 9 (Right). Figure 9 (right) clearly shows a difference in the ACF for each cluster, the ACFs indicate that clusters 2, 3, 4 and 5 exhibit similar model structure only off-set by differences in model parameters, while clusters 0 and 1 are distinctively different.

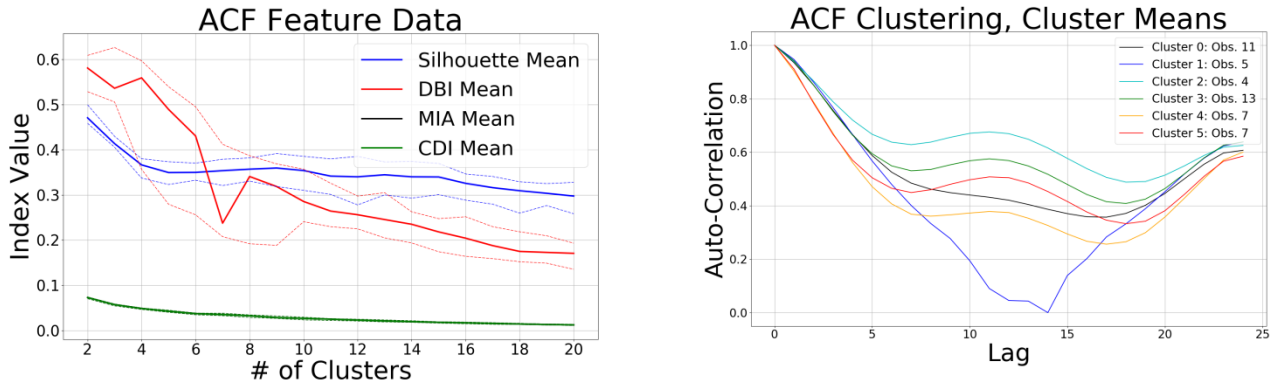


Figure 9 - (Left) Cluster validation index development, pseudo cross-validated, indicates a change at 7 clusters by the DBI index. The bold line is average index value while the dashed lines indicate minimum and maximum observed value for each index. (Right) Cluster means from each of the 6 cluster which have 4 or more members.

The cluster membership overlay in Figure 10 illustrates how the dispersion of the different classes has significantly reduced the overlap between clusters in the region lag 8 to lag 17. The clustering of ACF features result in clearly separable clusters.

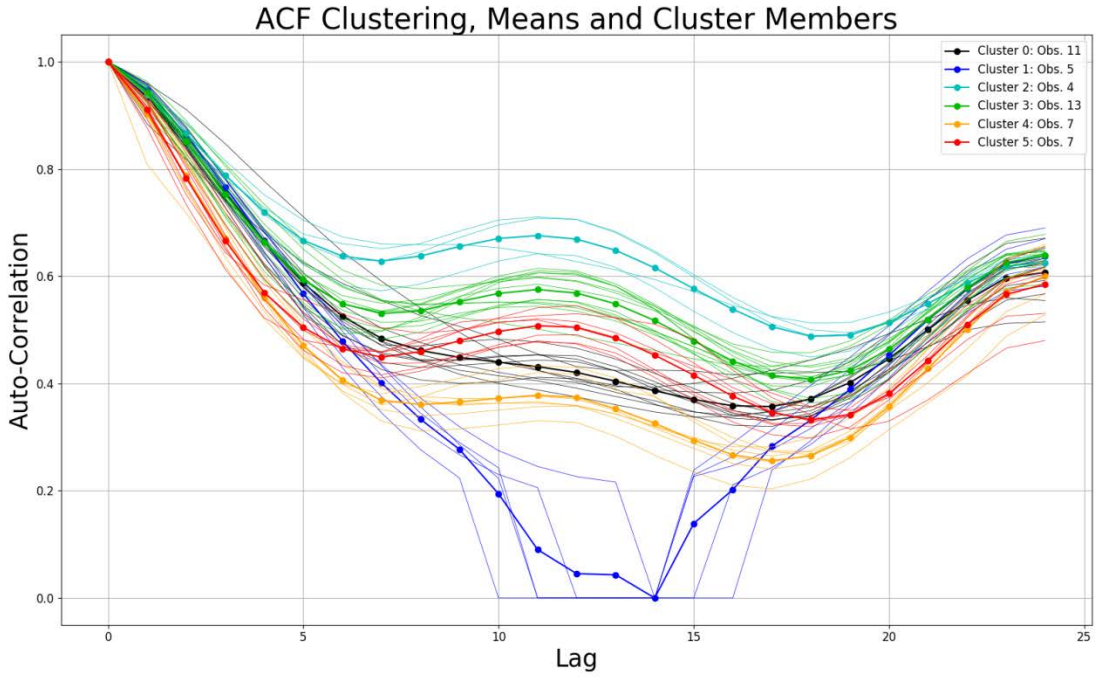


Figure 10 – Cluster member overlay. Colors represent cluster membership, and dots indicate cluster mean. It is clearly seen that the cluster dispersion and subsequently overlap between different clusters is very small in the region lag 8 to lag 17, yielding a better discriminatory power.

There are more clusters identified with the ACF clustering compared to scaled clustering, the resulting ACF clusters could potentially be sub-clusters of the scaled clusters. Table 5 shows this is not the case as the 7 ACF clusters are scattered throughout the 4 scaled clusters. This clearly indicates a difference in the resulting clustering by the 2 approaches.

Method		ACF							**Total**
	Clusters #	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**Total**
Normalized	**0**	7	3		1	3	2	2	18
	1				2		2		4
	2	2		2	1	4	3		12
	3	2	2	2	9				15
	Total	11	5	4	13	7	7	2	49

Table 5 - Cluster overlap table. Columns show the 7 clusters from the ACF transformation clustering, while rows show clustering with the Normalized. All Normalized clusters are scattered across several ACF clusters which shows that the detailed ACF clustering is not just a subset of the normalized clustering, but are entirely different clusters.

5.3 Cluster Performance: Wavelet Feature Extraction

It is no trivial task to select the wavelet best suited to fit a given signal. The collection of documented wavelet basis function is extensive. Only wavelets already implemented in the PyWt python package has been evaluated to keep the task feasible in this study. This was done by fitting each wavelet family to every HX station, and selecting the wavelet with best overall fit on all HX stations. The Coiflet 16 wavelet shown in Figure 11 had the best overall fit. The wavelet transformation of the original data was done by retaining only significant wavelet coefficients and creates a wavelet coefficients input data set for the K-Means.

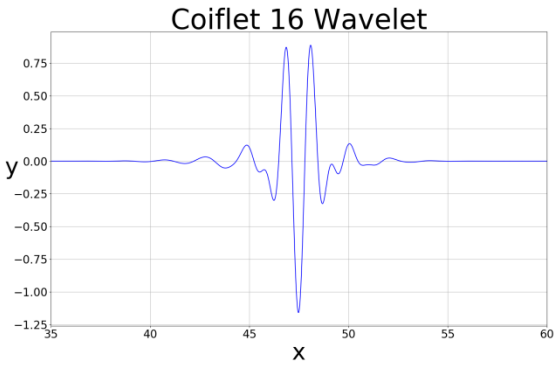
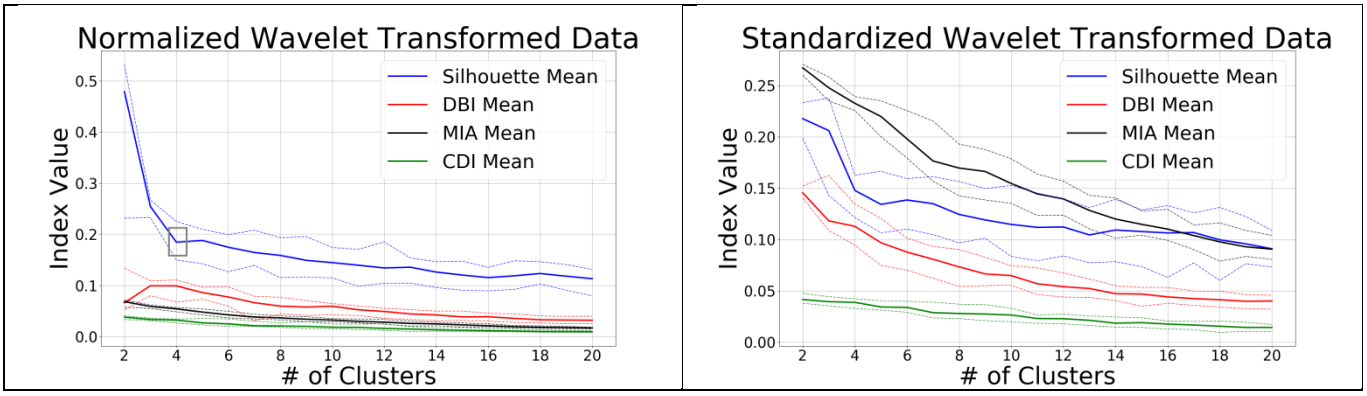


Figure 11 – Coiflet 16 mother wavelet selected, through testing, as the best wavelet for transformation of the smart meter data.

Finally we generate the cluster validation index progression shown in Figure 12 by applying the exact same procedure for estimating number of clusters as previously shown. All scaling, except mean-divided, indicate 4 clusters for the data. Mean-divided shows 4 or 5 clusters could be relevant. Normalizing and standardizing result in the most balanced clusters, but the index development for normalizing is more abrupt than for standardizing. All but the mean-divided scaling indicate 4 clusters. Even the original unscaled meter data suggests 4 clusters after wavelet transform. The 5 cluster solution for mean-divided results in very unbalanced clusters and is disregarded. We therefore conduct clustering on normalized data to ensure scaling is not an influencing factor to the wavelet coefficients for the remainder of the wavelet transform analysis.



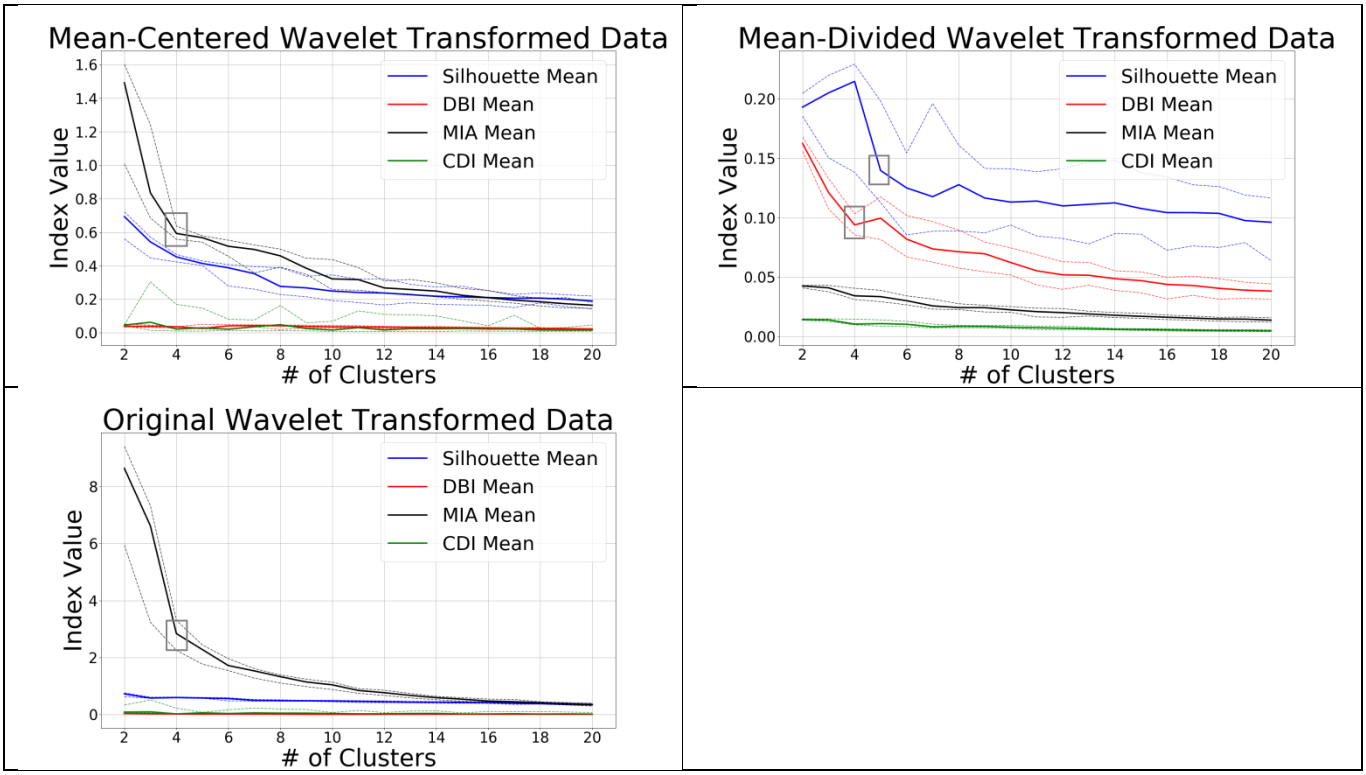


Figure 12 - All initial scaling, except mean-divided indicate same number of clusters for the wavelet transformed data. Coincidentally the same number of clusters suggested by the normalization scaling applied in 5.1. Mean-divided suggest 5 clusters which are more unbalanced and thus disregarded.

The resulting normalized cluster balance is (20, 4, 9, 16) which yield clusters with at least 4 members. Additionally, normalized scaling makes the results comparable to the clustering in section 5.1. The clustering using wavelet coefficients is shown in Figure 13. The plots are significantly different from the ACF and Normalized clustering due to how the coefficients are structured in the input data. Coefficients to the far right are from the lowest level of the pyramid and many are set to zero. From the figure it can be seen that the black clusters are shadowed by the red and green clusters, which was also observed with scaled clustering.

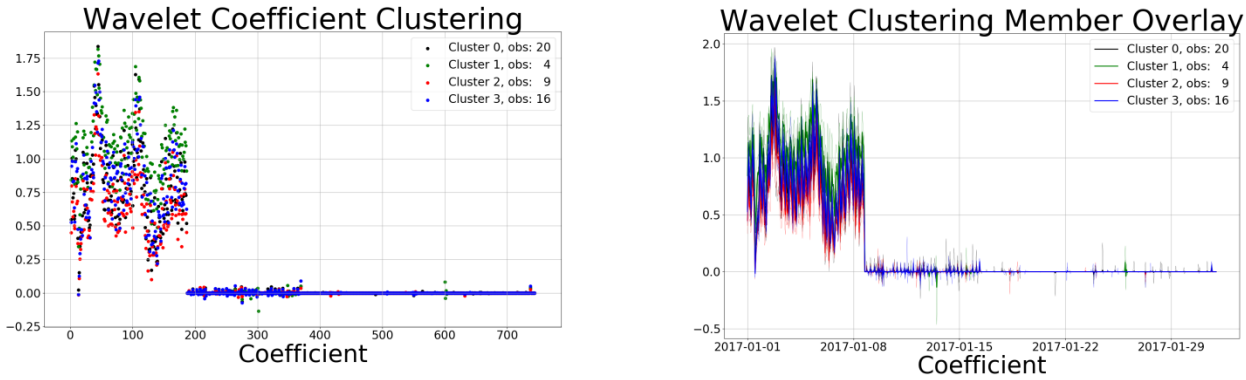


Figure 13 - (Left) Wavelet coefficient clustering. From left to right coefficients are from decreasing level of the pyramid algorithm. The number of non-significant coefficients increase towards the right side of the graph. (Right) cluster member overlay.

The resulting clustering overlap between Normalized and wavelet clustering is shown in Table 6. The clustering is very similar with only 6 of the 49 HX stations clustered differently between the two methods.

Method		Wavelet				
	Cluster #	0	1	2	3	Total
Normalized	0	16			2	18
	1		4			4
	2	3		9		12
	3	1			14	15
	Total	20	4	9	16	49

Table 6 - Clustering overlap table between Normalized and wavelet transformed data. Columns show wavelet transformed clustering and rows show Normalized clustering. Normalized and Wavelets produce similar clusters with this data set. Nearly all Normalized clusters are mapped 1:1 to the corresponding Wavelet cluster. The two methods yield similar clustering in this case.

5.4 Comparison of Clustering

The different clustering performed in sections 5.1, 5.2 and 5.3 have all utilized validation indices for selecting the number of clusters in the data, with "unknown ground truth" [33]. The three different methods applied in this paper results in different clustering solutions, though Normalization and wavelet transformation agree on 43 of the 49 HX stations in 4 different clusters. ACF finds 7 different clusters which are not a subset of the Normalization clustering but entirely different clusters. All three methods result in statistical significant groups at 5% confidence level, meaning that overall the clustering in all three cases result in some clusters that are statistically distinguishable. The statistical testing conducted was analysis of variance [40] implemented in the python package Scipy [41]. Caveat to this is that all three clustering result also produced shadow clusters, clusters that shadow other clusters resulting in clusters that statistically are impossible to distinguish. This phenomenon was observed in all three cases, but ACF had fewer shadow clusters than Normalization and Wavelets.

While we have shown the transformation of the input data enables K-Means to handle autocorrelated data, we are unable to show statistical difference between the three cases. Another measure for evaluating the clustering is analyzing the computational effort needed to perform the clustering. All three cases preprocess the input data, in processes that can be run in constant time and its influence on the overall runtime is negligible compared to K-Means lower bound runtime of $2^{\sqrt{n}}$ [42] and upper bound of $O(k^n)$ [42], where k is clusters and n is observations. The reduction of the input data via the wavelet or ACF feature extraction decrease from 744 to 161 or 24 coefficients respectively results in a dramatic decrease in minimum and worst case computational effort needed to cluster the data, Table 7.

	Scaled	ACF	Wavelet
Scaling / Transform	Constant time	Constant time	Constant time
Size of input data (n)	744 x 49	25 x 49	744 x 49
Best case running time	$2^{\sqrt{744}}$	$2^{\sqrt{24}}$	$2^{\sqrt{161}}$
Worst case running time	4^{744}	7^{24}	4^{161}

Table 7 - Clustering method comparison. The different scaling and transformations presented are able to identify clusters in the data. Selecting one instead of another is difficult. In this case we can see the worst-case running time for the ACF clustering is better than the scaled or wavelet transformed data.

6. Discussion

This paper has applied learnings from smart meter electricity consumption clustering to district heat exchange station clustering. The aim has been to investigate if the same methods are readily applicable to district heat clustering. The data utilized in this paper is generated by smart meters installed at Heat Exchange stations, the intersection where heat is transferred from the transmission grid to the distribution grid. We treated each HX station as a consumer and clustered according to their consumption pattern. The data in this study is kindly supplied by the district heating company AVA of Aarhus. 49 HX stations were included in the analysis. The study has focused on clustering hourly consumption data for the entire month of January 2017. AVA expects that clustering of smart meter data, in time, can help improve production efficiency by 1-2%, generating savings of 1.3-2.7 million euros a year. As mentioned in [2] clustering of electricity smart meter data has been performed at different time scales, from daily to a full year with hourly consumption recordings.

The most prevalent method for consumption clustering has been identified to be the K-Means algorithm, This paper has discussed the clustering performance and possible improvements to the algorithm. K-Means popularity among clustering algorithms can be attributed to its widespread implementation in popular analytical software, its stability and generally good performance. We used the python Sci-Kit Learning implementation of K-Means to generate the clustering, we scaled the input data to ensure pattern clustering rather than consumption volume clustering, and we applied four different cluster validation metrics for evaluating the optimum number of clusters. We found that K-Means clustering on scaled smart meter data to be academically viable, though the resulting clusters can be very unbalanced. The identified clusters exhibit large dispersion resulting in overlapping clusters, a phenomenon also encountered in electricity meter clustering papers such as [3] rendering the practical applicability of the clustering less feasible.

The clusters identified with ACF are not merely subsets of the scaled clustering but entirely different clusters. Furthermore we performed a wavelet basis transformation of the original data. The ACF and Wavelet transformation enabled K—Means to account for autocorrelation. The ensuing clustering improved separation of clusters, and it resulted in nearly identical clusters as scaling did.

A surprising result from the ACF clustering is that 5 out of 7 clusters have identical structure only offset by differences in coefficients values. This suggests that most of the HX stations consumption pattern can be described by the same underlying time series model, regardless of composition of the consumers supplied by the individual HX stations. This is surprising as the demographic and consumer composition of an area intuitively is expected to influence the consumption pattern, however the clustering using ACF indicates this is not the case. As it is the first time it has been encountered further research is required in this area. Deeper knowledge of the different areas served by the HX stations could help in understanding the patterns observed.

Further improvements to this study include investigation into the stability of the clusters. Are the clusters stabile over time- or do the cluster members transition between clusters? For a study of these potential

transitions and cluster stability over time, elimination of exoteric influences, such as; weather and temperature is needed.

Stability of the clustering was ensured by the concept of cross-validation. Cross-validation may be tweaked for unsupervised learning such as to give an indication of how much each cluster fluctuates. Though there is not a simple unified approach for cross-validation in an unsupervised setting the paper [29] discusses cross-validation for principal components selection. In this paper we have applied leave-one-out cross-validation, utilizing validation indices as pseudo response variables for cross-validating the estimate of the optimum number of clusters.

7. Conclusion

This paper presents novel ways of transforming smart meter input data before applying K-Means clustering. The transformation can be regarded as feature extraction with subsequent clustering of the features rather than the original meter data. With success we empirically confirm the existence of autocorrelation in the meter data and are able to cluster the heat exchange stations with regard to their autocorrelation function. The effect of clustering the autocorrelation enables the K-Means algorithm to account for the autocorrelation inherent in the meter data. To the best of the authors' knowledge this transformation has not been performed on electricity meter data or district heat meter data before. The resulting clustering when applying autocorrelation features generates more visually distinguishable clusters. However, we were unable to find significant differences in the clustering results. Unfortunately all cases contain shadow clusters that are statistically indistinguishable from neighboring clusters. Apart from showing that clever preprocessing of input data to K-Means can result in good cluster performance for the ACF transformation, the reduction in upper bound runtime is reduced from 4^{744} to 7^{24} which is a significant reduction in computational effort needed for the clustering process.

The main focus of this paper has been to apply the most prevalent clustering method from electricity consumption clustering to district heating data. The literature concerning district heat consumption clustering is still limited. This paper has proposed solutions to the K-Means algorithms limited ability to account for autocorrelation by transformation of input data. We conclude that the K-Means algorithm is indeed capable of clustering district heating consumption data. While diligently preprocessing the data we can further increase K-Means applicability on more complex data structures, in this case enable accounting for auto-correlation.

8. Executive Summary

- This paper clusters 49 district heat exchange stations applying smart meter consumption data, applying hourly recordings from entire January 2017, summing to 744 observations per station.

- K-Means was selected as clustering method due to its prevalence in the electricity consumption clustering literature.

  - Baseline clustering was done on normalized data, and validated using MIA, CDI, DBI and Silhouette index.
  - Autocorrelation was shown in the data and preprocessing of input enabled K-Means to account for autocorrelation information in the clustering.
    - Autocorrelation feature extraction and Wavelets were applied to account for autocorrelation.

- K-Means was able to cluster district heat smart meter data.

  - The clustering was improved by preprocessing the data prior to clustering.
    - Autocorrelation features enabled more detailed clusters
    - Wavelet features enabled compression of input data with identical clusters as baseline.

- Preprocessing of input data can enable K-Means to account for different inherent data structures, improving overall clustering and worse-case runtime. The autocorrelation feature extraction developed finer clusters drastically improved the run-time.

9. Acknowledgement

The authors wish to thank AffaldVarme Aarhus for participating in the research project, by making relevant data available. Pia Thomsen and Rikke Brinkø Berg for valuable input and endless proof reading. This work is part of the CITIES (Centre for IT-Intelligent Energy Systems in Cities) research project.
This work was supported by the Danish Innovation Found. Grant DSF 1305-00027B (Det Strategiske Forskningsråd). The Danish Innovation Found had no involvement in the study design.

Declaration of interest: None.

References

[1] Fjernvarme D. Fakta om fjernvarme 2017:1. http://www.danskfjernvarme.dk/presse/fakta-om-fjernvarme (accessed March 16, 2018).

[2] Tureczek. Alexander M, Nielsen. Per S. Structured Literature Review of Electricity Consumption Classification Using Smart Meter Data. Energies 2017;10:584. doi:10.3390/en10050584.

[3] Viegas JL, Vieira SM, Melício R, Mendes VMF, Sousa JMC. Classification of new electricity customers based on surveys and smart metering data Commission for Energy Regulation. Energy 2016;107:804–17. doi:10.1016/j.energy.2016.04.065.

[4] Chicco G, Napoli R, Piglione F, Postolache P, Scutariu M, Toader C. Load pattern-based classification of electricity customers. IEEE Trans Power Syst 2004;19:1232–9. doi:10.1109/TPWRS.2004.826810.

[5] Basu K, Debusschere V, Douzal-chouakria A, Bacha S. Time series distance-based methods for non-intrusive load monitoring in residential buildings. Energy Build 2015;96:109–17. doi:10.1016/j.enbuild.2015.03.021.

[6] Chicco G, Napoli R, Piglione F. Comparisons Among Clustering Techniques for Electricity Customer Classification. IEEE Trans Power Syst 2006;21:1–7.

[7] Mcloughlin F, Duffy A, Conlon M. A clustering approach to domestic electricity load profile characterisation using smart metering data. Appl Energy 2015;141:190–9. doi:10.1016/j.apenergy.2014.12.039.

[8] Räsänen T, Voukantsis D, Niska H, Karatzas K, Kolehmainen M. Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data. Appl Energy 2010;87:3538–45. doi:10.1016/j.apenergy.2010.05.015.

[9] Granell R, Axon CJ, Wallom DCH. Impacts of Raw Data Temporal Resolution Using Selected Clustering Methods on Residential Electricity Load Profiles. IEEE Trans Power Syst 2015;30:3217–24.

[10] Chicco G, Sumaili Akilimali J. Renyi entropy-based classification of daily electrical load patterns. IET Gener Transm Distrib 2010;4:736–45. doi:10.1049/iet-gtd.2009.0161.

[11] Park S, Ryu S, Choi Y, Kim J, Kim H. Data-Driven Baseline Estimation of Residential Buildings for Demand Response. Energies 2015;8:10239–59. doi:10.3390/en80910239.

[12] Tsekouras GJ, Hatziargyriou ND, Member S, Dialynas EN, Member S. Two-Stage Pattern Recognition of Load Curves for Classification of Electricity Customers 2007;22:1120–8.

[13] Carpaneto E, Chicco G, Napoli R, Scutariu M. Electricity customer classification using frequency – domain load pattern data 2006;28:13–20. doi:10.1016/j.ijepes.2005.08.017.

[14] Kang J, Lee J. Electricity Customer Clustering Following Experts' Principle for Demand Response Applications. Energies 2015;8:12242–65. doi:10.3390/en81012242.

[15] Haben S, Singleton C, Grindrod P. Analysis and Clustering of Residential Customers Energy Behavioral Demand Using Smart Meter Data. IEEE Trans Smart Grid 2015;7:136–44.

605 doi:10.1109/TSG.2015.2409786.

606 [16] Gouveia JP, Seixas J. Unraveling electricity consumption profiles in households through clusters :
607 Combining smart meters and door-to-door surveys. Energy Build 2016;116:666–76.
608 doi:10.1016/j.enbuild.2016.01.043.

609 [17] Kavousian A, Rajagopal R, Fischer M. Determinants of residential electricity consumption : Using smart
610 meter data to examine the effect of climate , building characteristics , appliance stock , and occupants ’
611 behavior. Energy 2013;55:184–94. doi:10.1016/j.energy.2013.03.086.

612 [18] Mcloughlin F, Duffy A, Conlon M. Characterising domestic electricity consumption patterns by dwelling
613 and occupant socio-economic variables : An Irish case study. Energy Build 2012;48:240–8.
614 doi:10.1016/j.enbuild.2012.01.037.

615 [19] Ndiaye D, Gabriel K. Principal component analysis of the electricity consumption in residential dwellings
616 2011;43:446–53. doi:10.1016/j.enbuild.2010.10.008.

617 [20] Ozawa A, Furusato R, Yoshida Y. Determining the relationship between a household ’ s lifestyle and its
618 electricity consumption in Japan by analyzing measured electric load profiles. Energy Build
619 2016;119:200–10. doi:10.1016/j.enbuild.2016.03.047.

620 [21] Gadd H, Werner S. Heat load patterns in district heating substations. Appl Energy 2013;108:176–83.
621 doi:10.1016/j.apenergy.2013.02.062.

622 [22] Gianniou P, Liu X, Heller A, Nielsen PS, Rode C. Clustering-based analysis for residential district heating
623 data. Energy Convers Manag 2018;165:840–50. doi:10.1016/j.enconman.2018.03.015.

624 [23] Cormen T, Rivest R, Leiserson C. Introduction to Algorithms. second edi. The MIT Press; 2001.

625 [24] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine
626 Learning in Python. J Mach Learn Res 2012;12:2825–30. doi:10.1007/s13398-014-0173-7.2.

627 [25] Lattin J, Carrol JD, Green PE. Analyzing Multivariate Data. vol. 46. 1. st. Duxbury; 2004.
628 doi:10.1198/tech.2004.s798.

629 [26] Bishop CM. Pattern Recognition and Machine Learning. 1st ed. Springer; 2006. doi:10.1117/1.2819119.

630 [27] López JJ, Aguado JA, Martín F, Mu F, Rodríguez A, Ruiz JE. Hopfield – K -Means clustering algorithm : A
631 proposal for the segmentation of electricity customers 2011;81:716–24.
632 doi:10.1016/j.epsr.2010.10.036.

633 [28] Maulik U, Bandyopadhyay S. Performance evaluation of some clustering algorithms and validity indices.
634 IEEE Trans Pattern Anal Mach Intell 2002;24:1650–4. doi:10.1109/TPAMI.2002.1114856.

635 [29] Perry PO. Cross-Validation for Unsupervised Learning 2009.

636 [30] Wasserman L. All of Statistics. vol. C. Springer; 2003. doi:10.1007/978-0-387-21736-9.

637 [31] Friedman J, Hastie T. The Elements of Statistical Learning. 1st ed. Springer; 2008.

[32] Madsen H. Time Series Analysis. 1st ed. Chapman & Hall/CRC; 2007.

[33] Warren Liao T. Clustering of time series data - A survey. Pattern Recognit 2005;38:1857–74. doi:10.1016/j.patcog.2005.01.025.

[34] Barford LA, Fazzio RS, Smith DR. An introduction to wavelets. Hewlett-Packard Labs, Bristol, UK, Tech Rep HPL-92-124 1992;2:1–29. doi:10.1109/99.388960.

[35] Morchen F. Time series feature extraction for data mining using DWT and DFT 2003:1–31. doi:citeulike-article-id:3973352.

[36] Li T, Li Q, Zhu S, Ogihara M. A survey on wavelet applications in data mining. ACM SIGKDD Explor Newsl 2002;4:49–68. doi:10.1145/772862.772870.

[37] Nason GP. Wavelet Methods in Statistics with R. 2008. doi:10.1007/978-0-387-78171-6.

[38] Graps A. An introduction to wavelets. IEEE Comput Sci Eng 1995;2:50–61. doi:10.1109/99.388960.

[39] Wasilewski F. PyWavelets 2006.

[40] Johnson RA. Probability and Statistics for Engineers. 6th ed. Miller & Freund's; 2000.

[41] Jones E, Oliphant T, Peterson P. SciPy: Open Source Scientific Tools for Python 2001.

[42] Arthur D, Vassilvitskii S. How slow is the k-means method? Proc Twenty-Second Annu Symp Comput Geom - SCG '06 2006:144. doi:10.1145/1137856.1137880.

Paper 4 – Stability of Electricity Smart Meter Consumption Clusters over Time

Stability of Electricity Smart Meter Consumption Clusters over Time

Alexander Martin Tureczek,

Systems Analysis, Management Engineering, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark; atur@dtu.dk

Abstract:

The last decade much research regarding electricity consumption clustering using smart-meter data has been conducted. The general consensus is a positive ability for the smart-meter electricity data to be applied for consumption clustering. Little research has been directed towards the generalizability of the clusters, i.e. are the clusters valid across time such that the clustering is independent of the time period selected? This paper applies hourly smart-meter electricity consumption readings to create weekly consumption clusters for more than 25,000 households throughout 2011. The weekly clustering allows for analysis of the development of clusters over the year. The clusters are created using the currently prevailing clustering techniques K-Means, but with the input data are preprocessed to manage autocorrelation. The paper develops a novel method for evaluating how members of clusters transition between clusters over the course of the year. Transition is categorized into two mapping types; 1:1 perfect stability, and 1:n total instability of clusters. The analysis shows that the clustering solutions are only valid the week they are created and that no cluster composition is repeated throughout the year. The findings show that the current prevailing methodology for smart-meter electricity clustering does not produce time stabile clustering solutions and thus hinders electric utilities to leverage the clustering.

Highlights:

Weekly clustering of more than 25,000 households for the entire 2011.

Methodology developed enabling analyses of cluster stability over time.

K-Means clustering of smart-meter electricity data produce unstable clusters.

Weekly smart-meter electricity consumption clusters are valid in current week only.

Keywords: K-Means, electricity clustering, consumption clustering, cluster stability, smart meter data

1. Introduction

The advent of modern digital electricity-metering systems, known as smart meters, boasts of endless possibilities regarding grid control and consumption flexibility through the in-depth recording of demand [1]–[3]. For more than a decade researchers have successfully analyzed these meter data to identify consumption patterns [4]. Numerous projects have applied K-Means and other clustering algorithms from machine learning to identify various consumption patterns hidden in the smart-meter data [5], [6]. Selecting the optimum number of consumption clusters using cluster validation indices, the latter evaluate the clusters and aid in the selection of optimum solutions. The indices evaluate both inter- and intra-cluster distances, selecting the most stable solution for the period being analyzed, but they do not supply any information about the stability of the solution across different time periods.

Researchers and private stakeholders are trying to produce consumption-clustering solutions applicable outside academia to facilitate value propositions for both utilities and consumers. Some utilities develop mobile applications allowing customers to audit and analyze their consumption on a daily basis [7], while research papers propose solutions to aid utilities in optimizing their production and to identify consumption flexibility for advanced tariff schemes and incentives[5], [8], [9]. It has been suggested that the cluster solutions that are achieved through the current state of the art in smart-meter consumption clustering exhibit too much within-cluster variance to be able to create viable and uniquely identifiable clusters[4], [10], [11].

The stability of the clustering that is performed is seldom investigated beyond the validation indices, which describe the clustering for the current period. For clusters to be truly applicable beyond academia, they need to be defined in such a way that they are meaningful and persistent. Therefore, the stability of the clusters across time periods must be investigated to ensure that cluster solutions remain the same and that the transition between clusters is quantified.

This raises two questions. First, are the clusters independent of the time of year at which the data have been recorded? Secondly, are meters that were clustered together in January also clustered together in February etc., or do they transition individually between clusters across time periods, resulting in unstable cluster definitions? Few papers have investigated the stability of smart-meter consumption-clustering solutions across time [12]. The hypothesis of this paper is that the clustering solutions that currently represent the state of the art in smart-meter analyses are stable across time periods.

To investigate this hypothesis, the present paper presents the clustering of over 25,000 households on a weekly basis for the whole of 2011. The year will be split into its four quarters, which are closely tied to the seasons, in order to evaluate and quantify cluster stability through the course of a year, but evaluating each quarter independently, as the weeks within them are expected to exhibit similar levels of consumption. Electricity consumption data have been collected throughout 2011. In the Danish energy system, 64% of household heating is supplied by large district-heating utilities and not by electricity [13]. All the meters included in this analysis are connected to district heating and thus do not use electricity for heating. To evaluate the stability of clustering across weeks, this paper develops a novel method for evaluating whether there is a transition between clusters between weeks.

The remainder of this paper is divided into seven sections. Section 2 presents a review of the literature on energy-consumption clustering, followed by a description of the data analyzed in the present paper in section 3. The methodology followed is described in section 4, which includes brief descriptions of K-Means (4.1), Cluster Validation Indices (4.2) and Autocorrelation Features (4.4). The section also includes a presentation of cross-validation for unsupervised learning (4.3) and introduces Varatio, a novel method of evaluating the cluster transition through variance evaluation. Section 5 presents the results, with discussion in section 6, followed by a conclusion in section 7.

2. Literature Review

Paper[4] investigated the current state of the art in smart-meter energy-consumption clustering. The K-Means clustering algorithm and derived methods such as fuzzy K-Means are identified as the most prevalent clustering algorithms in smart-meter energy consumption-clustering [4]. In the literature it is used solely as either the clustering method [14] or the benchmark for testing more advanced methods [15]. More than ten clustering methods were identified, the runners-up being Hierarchical Clustering and Follow the Leader. The K-Means algorithm is unable to include intrinsic information in the clustering, [10], [11] revealed the existence of autocorrelation in smart-meter electricity and district heating data, which by default K-Means is unable to include in the clustering.

To validate the clustering performed by unsupervised clustering methods, numerous cluster validation indices have been developed, more than fourteen different indices being identified [1]. Popular indices are the Cluster Dispersion Index [5], the Davies-Bouldin Index [16], Mean Index Adequacy [17] and the Silhouette Index [18]. These indices will also be used in this paper. Apart from the cluster validation indices, few papers are concerned with the stability of the clusters over time [12].

Papers [10], [11] introduce methods for conducting the cross-validation of unsupervised clustering using smart-meter data. There is no general framework for cross-validation in unsupervised learning [19]. Only two papers were found applying cross-validation to smart-meter data, and both did so in a supervised setting [20], [21]. This paper will apply the pseudo-cross-validation methodology presented in [10].

Throughout the literature, there is no consensus regarding either recording frequency or how many smart-meter recordings to include in the clustering. With respect to recording frequency, a single paper used second resolution [17], while most used frequencies of fifteen minutes [5], [9], [16] to sixty minutes [14], [22], [23]. The number of consecutive recordings applied for clustering is anywhere from one day [24], [25] to several years [9], [14], [15]. This paper will use recordings at sixty-minute intervals and weekly time periods.

3. Data Description and Preparation

This paper analyses a subset of the SydEnergi data set introduced in [10]. The final size of the SydEnergi electricity-consumption data set analyzed in this paper covers 26,562 households, (semi-)detached houses and apartments in the four post codes of the city of Esbjerg all connected to the municipal district-heating grid. As

the two dwelling types shown in [10] exhibit similar consumption structures, we include them both. The data preprocessing stage was identical to that in [10] except for the exclusion of week-on-week meters and only including meters with readings for weeks 1 to 51 of 2011. As a result, the data displayed autocorrelation. Essential data set information is listed in Table 1.

Essential Data Information	Value
Country	Denmark
Region	Region Syd (Region South), post codes: 6700, 6705, 6710, 6715. Comprising the city of Esbjerg.
Supplier	SydEnergi Electric Utility Company
Initial size	26562 smart meters
Clear reduction	Data were cleaned prior to the analysis removing meters which have missing values or zero mean, median and variance consumption.
Missing values	No missing values observed.
Final size	26562
Recording frequency	60 min (aggregated from 15 mn)
Start	January 3rd 2011, week 1.
End	December 25th 2011, week 51.
Length	8568
Type	Apartments and (semi-)detached houses heated by district heating
Referral	Data referenced in [10]

Table 1. Essential data set information generating an overview of the analyzed data and the preprocessing undertaken prior to analysis.

4. Methodology

This section describes the methodology applied in this paper to cluster smart-meter consumption data. The K-Mean algorithm is briefly described in section 4.1. Section 4.2 describes the cluster validation indices employed when selecting the optimum number of clusters. In section 4.3, cross-validation of unsupervised clustering is described, and in section 4.4 an autocorrelation feature extraction is presented. Section 4.5 introduces Varatio, a novel measure of fit applicable when evaluating the stability of clusters across time periods.

4.1.K-Means

The literature study in section 2 found the family of K-Means-derived clustering algorithms to be the most prevalent in smart-meter consumption clustering. K-Means has therefore been selected as the clustering method that will be used in this paper for the analysis of cluster stability across time periods.

K-Means is a simple and efficient algorithm for clustering numeric data. As described in [10] and [11], the method is readily available in commercial and open-source software and is simple to implement if absent. The simplicity of the method introduces some caveats relating to its inability to handle inherent structures like

autocorrelation in smart-meter data, a drawback that has been addressed in a smart-meter data setting in [10], [11], describing possible ways of remedying the problem. One proposed method is autocorrelation feature extraction, that is, transforming the smart-meter data so that the transformed data explicitly include information on autocorrelation. We apply autocorrelation features to the data, thus enabling K-Means to account for the autocorrelation information. For a discussion of how to enable K-Means to handle autocorrelation in smart meter data, see [10], [11].

4.2.Cluster Validation Indices

With K-Means, the process of selecting the optimum number of clusters entails the application of cluster validation indices (CVI). These validation indices assess the performance of different clustering solutions by evaluating the different properties of the clusters. Some indices assess internal cluster dispersion, comparing it to the dispersion between clusters or average cluster distances. The intuition behind the indices is similar to error minimization in a supervised setting in that the indices are designed by construction such that they minimize or maximize a certain property of the clusters. This property is then evaluated for different numbers of clusters, ultimately striking a balance between optimization and parsimony: for example, fewer clusters if they exhibit resembling index values. An important point is that no index can calculate the true clusters, only being able to evaluate the clusters generated by the clustering algorithm. It is advisable to apply several indices simultaneously, as each index contributes to the selection, and specific data properties might render some indices ineffective.

This paper uses four indices selected due to the freequence of their deployment in the literature: Mean Index Adequacy (MIA), the Davies-Bouldin Index (DBI), the Cluster Dispersion Index (CDI) and the Silhouette Index. The individual properties of the indices are outlined in Table 2.

Index	Mathematical description	Properties
MIA	$$\boldsymbol{MIA} = \sqrt{\frac{1}{K}\sum_{k=1}^{K} \boldsymbol{d^2(C_k)}},$$	Average distance within class-member to class centroid, summarized across all classes. k is the number of clusters; $d^2(C_k)$ is the squared average distance within cluster k. A high MIA indicates large distances within the classes, e.g., large dispersion.
Cluster Dispersion Index (CDI)	$$\boldsymbol{CDI} = \frac{1}{\boldsymbol{d(C)}}\sqrt{\frac{1}{K}\sum_{k=1}^{K} \boldsymbol{d^2(C_k)}}$$	CDI prefers long inter-cluster distances and short intra-cluster distances [5]. Low values indicate good clustering. $d^2(C_k)$ is the squared average distance within cluster k, while $d(C)$ is the average cluster distance in the data [5].
Davies-Boudin Index (DBI)	$$\boldsymbol{DBI} = \frac{1}{K}\sum_{i=1}^{K} \boldsymbol{max}_{j\neq i} \frac{\boldsymbol{diam(C_i)} + \boldsymbol{diam(C_j)}}{\boldsymbol{d(C_i, C_j)}}$$	$diam(C_k)$ is the average diameter of a cluster, and $d(C_i, C_j)$ is the distance between cluster centers. K is the number of clusters. DBI correlates the mean distance of each class with the distance to the closest class [26]. Lower values of DBI implies that clustering algorithm has separated the data set properly [9].
Silhouette Index	$$\boldsymbol{Silhouette} = \frac{\boldsymbol{c'(x)} - \boldsymbol{c(x)}}{\boldsymbol{max}\{\boldsymbol{c(x)}, \boldsymbol{c'(x)}\}}$$ $$\boldsymbol{c'(x)} = \boldsymbol{\min_{y\in C'}} \boldsymbol{d(x,y)}$$	$c(x)$ is the average distance between vector x and all other vectors of the cluster c to which x belongs. $c'(x)$ is the minimum distance between vector x and all other vectors in cluster $\forall\ C' \neq C$ [14]. SI is between [–1, 1]; a higher number is better. A negative number indicates miss-clustering.

Table 2. Cluster validation indices applied to clustering smart meter data. The individual index amplifies certain properties of the clusters in an effort to distinguish and select the optimum clustering. As introduced in [11].

4.3.Cross-Validation

This paper will apply pseudo-cross-validation, using cluster validation indices as response variables, as introduced in [10], [11]. The process calculates the variability of the cluster validation indices through a ten-fold cross-validation of the times series. In effect this quantifies the fluctuation of the cluster validation index for each combination of clusters, quantifying the clustering performance of a given set of clusters. For each cluster combination, the pseudo-cross-validation calculates maximum, minimum and average index values, making no assumption about the underlying distribution. The maximum and minimum values are not interpretable as true confidence intervals. Although the maximum and minimum calculations are not statistical metrics for confidence intervals, the interval spread enables us to evaluate the fluctuation and thus the performance of each possible solution. The concept of cross-validation for unsupervised learning was discussed in [19] and proposed for smart-meter data in [10].

4.4.Autocorrelation Features

The data set analyzed in this paper was shown to contain autocorrelation, quantified using autocorrelation features [10] for improved clustering results. In time series analyses, autocorrelation quantifies the time dependence, that is, the influence of previous observations on the present observation. The autocorrelation features (ACF) enable the K-Means to incorporate autocorrelation information into the clustering solution. The ACF is invariant to consumption volumes, and thus the normalization usually required by K-Means makes no difference in this setting. Furthermore, the ACF acts as a dimensionality reduction by retaining only significant features, thus significantly reducing the dimension. In papers [10], [11] 24 features were retained for clustering. The features are equivalent to the 24 information lags in the autocorrelation function.

4.5.Varatio

Even though K-Means can produce divergent clustering results, the multiple rerun of the algorithm should eventually produce the best clustering solution, assuming sufficient random initializations. Consequently, clusters identified in January should be identified in the subsequent months, assuming they are stable.

For clusters created from smart-meter consumption data to be stable, there must be invariance in the period analyzed. If there is no such invariance, then it is impossible to achieve practical applicability of the resulting clusters, as they are only valid provisionally. As consumption does change with the seasons, the latter are defined as stable periods, though allowing for consumers to transition clusters between seasons.

Measuring the stability of clusters per season entails that the cluster solution created in week 1 of January are identifiable and identical in weeks 2-12 of the same season. That the solutions are identifiable in the number of clusters estimated using cluster validation indices and identical in that the same meters are clustered together across all weeks.

There are two possible outcome mappings when comparing clusters across weeks, illustrated in Figure 1: 1:1 and 1:n mapping. 1:1 mapping produces clusters that are identical across weeks, resulting in all meters

belonging to the same cluster continuing to fall into the same clusters in every clustered week. 1:n mapping scatters the members – that is, the individual clusters – across multiple clusters.

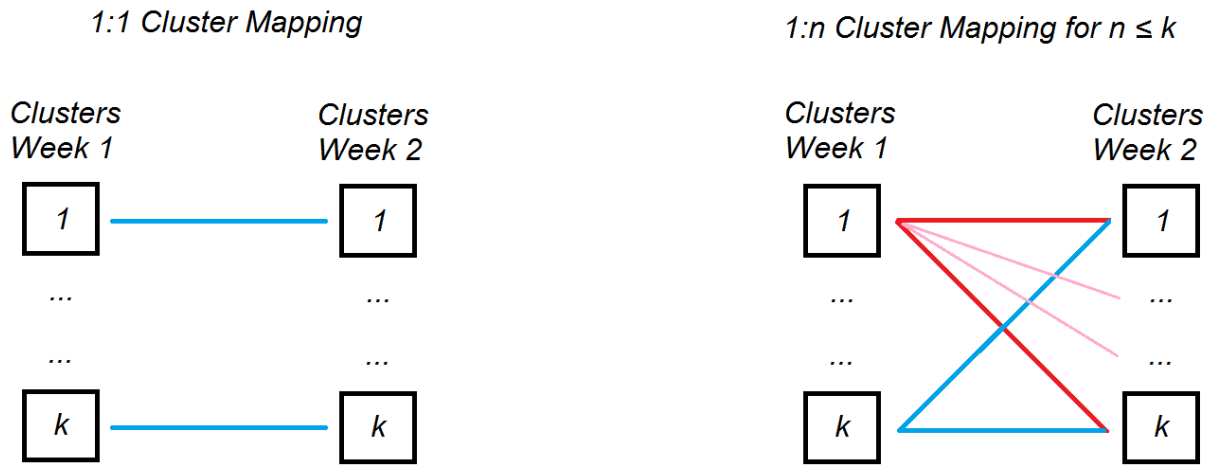


Figure 1. Cluster mapping types. 1:1 mapping delivers perfect mapping of meters from one cluster to another across time. 1:n mapping scatters the meters collected in one cluster to multiple clusters across time.

If two weeks appearing in the same season, say weeks 1 and 3 of January, it is rapidly established whether the cluster solutions in each week recommend the same number of clusters. If they do not, there is a discrepancy in the solution, and the clusters are unstable within the season. If the cluster solution recommends the same number of clusters, one question is whether the meters are clustered together across weeks in 1:1 or 1:n mapping.

Assuming there are five clusters in both weeks, then with 1:1 mapping the clusters retain their size, while 1:n mapping can result in size differences, though the same size could be an outcome. The Varatio introduces variance as a measure for establishing whether the meters are mapped using 1:1 or 1:n. It does this by realizing that, if there are five clusters, the worst-case scenario is that the meters are scattered uniformly across all five clusters in the following week's clustering. In this case the variance in the mapping will be zero or close to zero depending on the modulo of the clusters. The best-case scenario happens when the mapping is 1:1 and the meters are assembled into one of the five clusters, thus maximizing the variance of the mapping. Prior to the second clustering, the maximum attainable variance is calculated for each cluster by creating a vector with the length of the number of clusters, assigning zero to all elements, but one in which the number of meters of the investigated cluster is specified. This maximum variance is applied as the denominator of the Varatio, while the actual variance in the second clustering represents the nominator producing a ratio between the realized and the maximum variance.

$$Varatio = \frac{observed\ variance\ of\ distribution\ of\ second\ clustering}{variance\ of\ individual\ clusters\ in\ first\ clustering} * \% \qquad (1)$$

Varatio eliminates the need to investigate how the cluster is scattered across subsequent clustering by creating a measure for closeness to the 1:1 mapping. In stability analyses of the clusters, the distribution of the 1:n mapping is relevant to conveying information about the closeness to 1:1 mapping. 1:1 mapping using Varatio represents a hundred percent overlap of the maximum and observed variance, while 1:n mapping shows 0%. As Varatio does not provide guidelines for distinguishing the mapping evaluations it calculates, it is at the analyst's discretion to assess whether the mapping is applicable in the current setting. Figure 2 shows how Varatio is calculated in three distinct cases: best-case scenario, worst-case scenario and random mapping.

Figure 2. Two weeks each, with five clusters and fifty members in each cluster, are shown subjected to the Varatio calculations. Best-case scenario mapping is shown in the red box, with clusters in week 1 being mapped 1:1 into clusters in week 2, resulting in a Varatio of 100%. Worst-case scenario 1:n mapping is shown in the green box, representing the mapping of the clusters from week 1 uniformly across the clusters in week 2, resulting in a Varatio of 0%. Finally, Random Mapping shows how Varatio develops for different mappings from 1:1 to 1:n (approximately).

5. Results

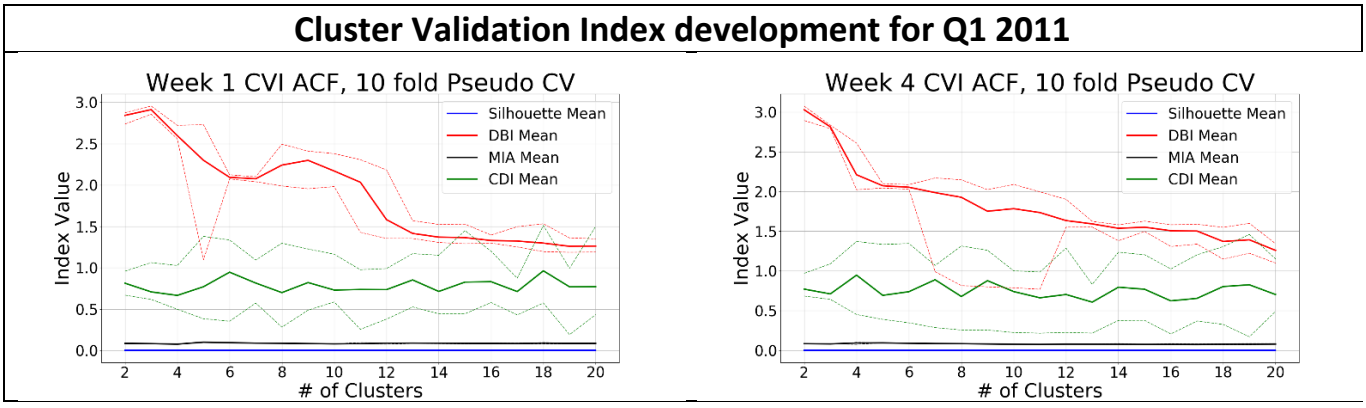
We analyze the SydEnergi electricity consumption data split into four different quarters of the year, representing a standardized division of the year roughly encompassing four different seasons. Weeks split between quarters are allocated by majority vote to the quarter that most days belong to. In the case of q1 and q2 the boundary is in week 13, with four days in q1 and three days in q2, allocating week 13 to q1. The quarters are then defined as follows: winter, weeks 1-13; spring, weeks 14-26; summer, weeks 27-39. fall, weeks 40-51.

For each week in each quarter the cluster validation indices are calculated using pseudo-cross-validation to select the optimum number of clusters. The analytical process is depicted in Figure 3, showing initial data preparation, followed by autocorrelation feature extraction, with subsequent clustering using K-Means. Selection of the optimum clustering is aided by the cluster validation indices (CVI). Finally Varatio is employed to evaluate the mapping of the clusters. The analysis is conducted for each week in line with the season. The individual seasonal results are presented below in sections 5.1 to 5.4.

Figure 3. Analytical process followed during analysis of the smart-meter data.

5.1.Stability of Clusters in Quarter 1

The cluster validation indices are calculated for all weeks with between two and twenty clusters in each week. Calculations for cluster validation indices (CVI) apply the pseudo-cross-validation. The CVI developments are shown for the selected four weeks of quarter 1 in Figure 4. The weeks all show different developments and thus different estimated optimum numbers of clusters, indicating that Q1 cluster members are mapping 1:n across the weeks. For every week, we select the optimum number of clusters that results in the fewest clusters. Over the course of the thirteen weeks in Q1, it can be argued that from weeks 4 to 13 the optimum number of clusters slowly progresses from four to six clusters. The estimated optimum number of clusters for each week are listed in Table 3.



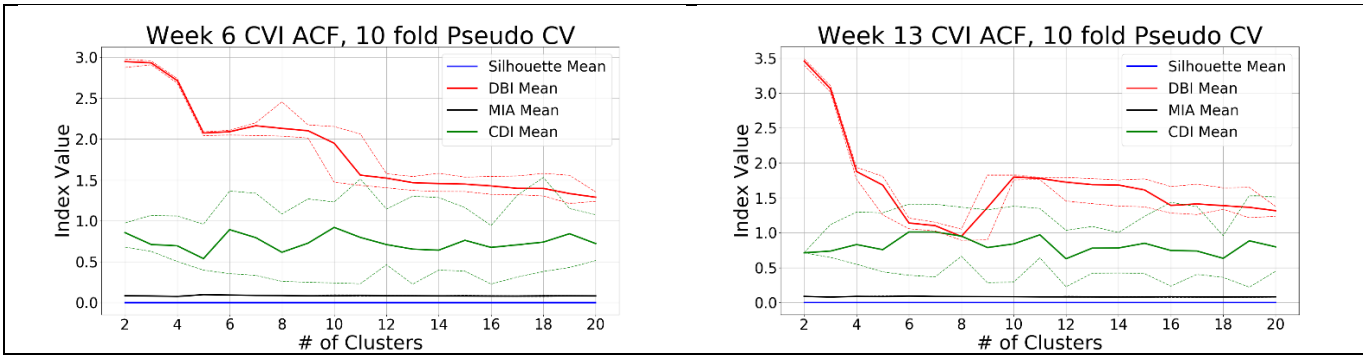


Figure 4. CVI development for four different weeks in Q1. The Silhouette and MIA indices are unable to compute applicable measures for cluster selection. CDI exhibits large variance. DBI is able to identify the optimum number of clusters. From the weeks it can be seen that multiple cluster numbers can be argued as being optimum. We select the solution that demands the fewest clusters on a weekly basis.

Q1 2011	
Week	**CVI Estimated Clusters**
1	12
2	12
3	12
4	4
5	5
6	5
7	4
8	4
9	6
10	5
11	6
12	6
13	6

Table 3. CVI-estimated optimum number of clusters for each week. A progression from four to six clusters is noticeable from weeks 4 to 13.

There is an indication that weeks 4 to 13 could consist of similar clusters, while weeks 1, 2 and 3 contrasts distinctly. Using Varatio, we investigate whether the mapping in weeks 1 to 13 is approximately 1:1. From Table 3 different weeks can be seen, suggesting a different number of clusters. We set the number of clusters each week in Q1 at six and apply Varatio to compare the mapping between the clusters. Table 4, Table 5 and Table 6 each represent a week in Q1 with different numbers of optimum clusters, but for comparability all tables show results for six clusters. The cell colorings of the tables indicate the Varatio value: dark green >50%, light green 20-50%, yellow 10-20%, light red 5-10% and dark red <5%. Few clusters in each week exceed 20% Varatio, indicating that the stability of clusters is at best 20% of the achievable stability, most results being far below 20%. All three tables indicate a 1:n mapping, which implies that the clustering is neither stable nor independent of the selected period that was clustered, even within a homogeneous period such as a quarter.

Week 4. Varatio overlap with rest of weeks in Q1

Overlap	4 to 1	4 to 2	4 to 3	4 to 4	4 to 5	4 to 6	4 to 7	4 to 8	4 to 9	4 to 10	4 to 11	4 to 12	4 to 13
Cluster 0	15%	17%	19%	100%	18%	19%	18%	20%	20%	20%	21%	20%	22%
Cluster 1	11%	11%	11%	100%	10%	10%	8%	10%	9%	9%	9%	8%	9%
Cluster 2	19%	18%	17%	100%	16%	14%	11%	13%	11%	11%	9%	10%	8%
Cluster 3	3%	4%	3%	100%	5%	3%	3%	3%	2%	3%	3%	1%	3%
Cluster 4	8%	9%	8%	100%	8%	8%	7%	9%	8%	10%	9%	8%	9%
Cluster 5	7%	7%	7%	100%	7%	7%	6%	7%	6%	8%	7%	6%	7%

Table 4. Week 4 clustering overlap with weeks in quarter 1. The percentages represent the percentage -attained variances of the maximum variance as defined by Varatio. Dark green indicates greater than 50% overlap, light green a 20-50% overlap, yellow a 10-20% overlap, light red a 5-10% overlap and dark red a <5% overlap.

Week 5. Varatio overlap with rest of weeks in Q1

Overlap	5 to 1	5 to 2	5 to 3	5 to 4	5 to 5	5 to 6	5 to 7	5 to 8	5 to 9	5 to 10	5 to 11	5 to 12	5 to 13
Cluster 0	7%	8%	8%	9%	100%	8%	7%	9%	9%	10%	9%	8%	10%
Cluster 1	6%	7%	6%	6%	100%	7%	6%	6%	6%	8%	7%	6%	7%
Cluster 2	19%	17%	15%	14%	100%	14%	10%	13%	11%	10%	10%	9%	7%
Cluster 3	12%	11%	11%	11%	100%	11%	8%	10%	9%	9%	9%	8%	9%
Cluster 4	3%	3%	3%	4%	100%	4%	3%	2%	3%	2%	3%	1%	3%
Cluster 5	16%	17%	19%	20%	100%	21%	18%	21%	22%	21%	22%	21%	23%

Table 5. Week 5 clustering overlap with weeks in quarter 1. The percentages represent the percentage -attained variances of the maximum variance as defined by Varatio. Dark green indicates greater than 50% overlap, light green a 20-50% overlap, yellow a 10-20% overlap, light red a 5-10% overlap and dark red a <5% overlap.

Week 9. Varatio overlap with rest of weeks in Q1

Week Overlap	9 to 1	9 to 2	9 to 3	9 to 4	9 to 5	9 to 6	9 to 7	9 to 8	9 to 9	9 to 10	9 to 11	9 to 12	9 to 13
Cluster 0	7%	8%	7%	7%	8%	7%	7%	8%	100%	10%	9%	7%	9%
Cluster 1	18%	17%	16%	13%	16%	15%	13%	16%	100%	13%	12%	12%	9%
Cluster 2	6%	6%	6%	6%	7%	7%	6%	6%	100%	8%	7%	6%	7%
Cluster 3	11%	10%	10%	11%	10%	10%	8%	11%	100%	10%	10%	9%	9%
Cluster 4	2%	3%	2%	3%	4%	3%	3%	4%	100%	3%	3%	3%	3%
Cluster 5	13%	14%	16%	16%	16%	16%	17%	19%	100%	19%	21%	20%	22%

Table 6. Week 5 clustering overlap with weeks in quarter 1. The percentages represent the percentage -attained variances of the maximum variance as defined by Varatio. Dark green indicates greater than 50% overlap, light green a 20-50% overlap, yellow a 10-20% overlap, light red a 5-10% overlap and dark red a <5% overlap.

We analyze quarters 2, 3 and 4 using an identical approach to that used in the analysis of quarter 1 and present the four most distinct CVI developments for each quarter, along with the CVI-estimated optimum table and three random overlap tables for each quarter.

5.2.Stability of Clusters in Quarter 2

Quarter 2 presents more homogeneous estimates of the optimum number of clusters across all weeks. The four CVI development graphs in Figure 5 indicate some differences in shape, but with agreement on six clusters as optimum for most weeks. The exact optimum number of clusters per week is shown in Table 7. The cluster Varatio overlap coefficients presented for three weeks in Table 8 indicate approximate 1:n mapping for Q2. This means the clusters are to a large degree not stable across the quarter.

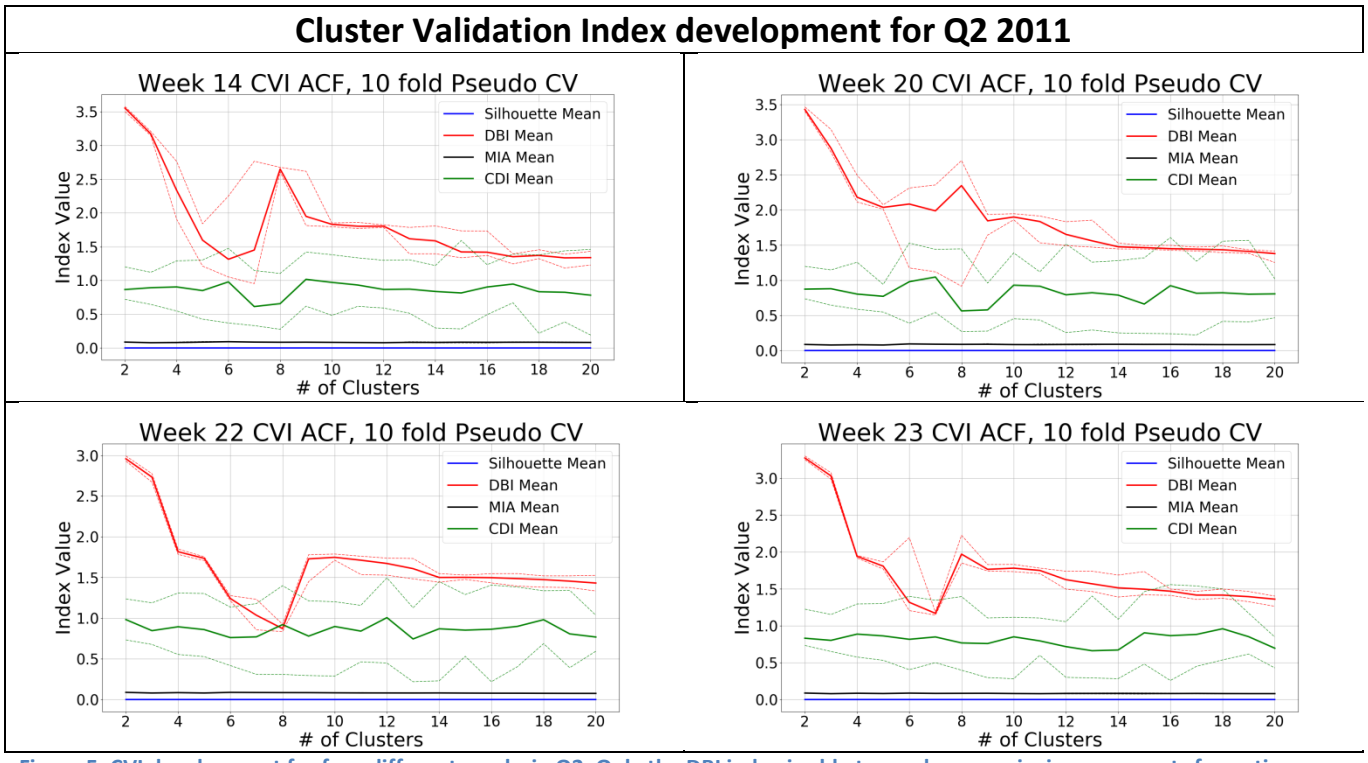


Figure 5. CVI development for four different weeks in Q2. Only the DBI index is able to produce convincing arguments for optimum cluster selection. The four weeks are selected based on their different estimates of optimum clusters.

Q2 2011	
Week	**CVI Estimated Clusters**
14	6

15	6
16	6
17	6
18	6
19	6
20	4
21	6
22	8
23	7
24	6
25	6
26	8

Table 7. CVI-estimated optimum number of clusters for each week in Q2. There is a more stable estimate of six clusters throughout the quarter, with few exceptions. Overall the CVI estimate indicates a more homogeneous quarter.

Week 19 Varatio overlap with rest of weeks in Q2

Week	19 to 14	19 to 15	19 to 16	19 to 17	19 to 18	19 to 19	19 to 20	19 to 21	19 to 22	19 to 23	19 to 24	19 to 25	19 to 26
Cluster 0	8%	9%	8%	9%	9%	100%	8%	7%	9%	7%	10%	8%	11%
Cluster 1	6%	6%	7%	7%	6%	100%	8%	7%	7%	5%	8%	6%	9%
Cluster 2	14%	13%	7%	11%	13%	100%	13%	15%	10%	13%	12%	13%	12%
Cluster 3	17%	19%	16%	19%	22%	100%	19%	20%	19%	18%	18%	17%	19%
Cluster 4	4%	4%	4%	5%	4%	100%	3%	5%	5%	4%	6%	3%	4%
Cluster 5	9%	9%	8%	9%	10%	100%	9%	8%	9%	9%	10%	9%	10%

Week 20 Varatio overlap with rest of weeks in Q2

Week	20 to 14	20 to 15	20 to 16	20 to 17	20 to 18	20 to 19	20 to 20	20 to 21	20 to 22	20 to 23	20 to 24	20 to 25	20 to 26
Cluster 0	11%	11%	6%	9%	10%	11%	100%	12%	9%	11%	11%	12%	11%
Cluster 1	19%	19%	17%	20%	22%	23%	100%	20%	20%	18%	19%	18%	20%
Cluster 2	6%	6%	6%	7%	5%	7%	100%	6%	7%	5%	8%	6%	9%
Cluster 3	9%	10%	9%	10%	11%	10%	100%	8%	10%	10%	11%	9%	11%
Cluster 4	3%	2%	2%	5%	3%	3%	100%	5%	4%	2%	4%	2%	4%
Cluster 5	8%	9%	9%	9%	9%	9%	100%	8%	9%	7%	10%	8%	11%

Week 22 Varatio overlap with rest of weeks in Q2

Week	22 to 14	22 to 15	22 to 16	22 to 17	22 to 18	22 to 19	22 to 20	22 to 21	22 to 22	22 to 23	22 to 24	22 to 25	22 to 26
Cluster 0	6%	6%	6%	6%	5%	6%	7%	5%	100%	5%	8%	5%	8%
Cluster 1	10%	10%	10%	10%	11%	11%	11%	8%	100%	10%	11%	11%	11%
Cluster 2	16%	17%	15%	17%	20%	20%	17%	17%	100%	17%	18%	16%	18%
Cluster 3	8%	8%	8%	9%	8%	8%	8%	7%	100%	7%	10%	7%	11%
Cluster 4	4%	4%	3%	4%	3%	4%	4%	4%	100%	4%	7%	4%	5%
Cluster 5	12%	11%	8%	10%	10%	11%	13%	13%	100%	14%	13%	13%	12%

Table 8. Varatio coefficients for each cluster combination in weeks 19, 20 and 22. Dark green indicates a 50%+ Varatio coefficient. Light green indicates Varatio estimated at between 20-50%, yellow at 10-20%, light red at 5-10% and dark red at <5% of maximum variance as defined by Varatio. In all three selected weeks of Q4, the mapping is approximately 1:n.

5.3.Stability of Clusters in Quarter 3

Quarter 3 is the period with the largest evidence of clustering, where the optimum could show ambiguity by presenting two or more relevant optima per week. Figure 6 presents four different CVI developments where there could be several optima. In the case of multiple optima in which six clusters were reasonable by CVI estimates, six clusters were selected, as shown in Table 9. As with the previous quarters, the Varatio coefficients in Table 10 reveal, that the mapping is approximately 1:n, implying that the clusters in Q3 are not stable across the quarter.

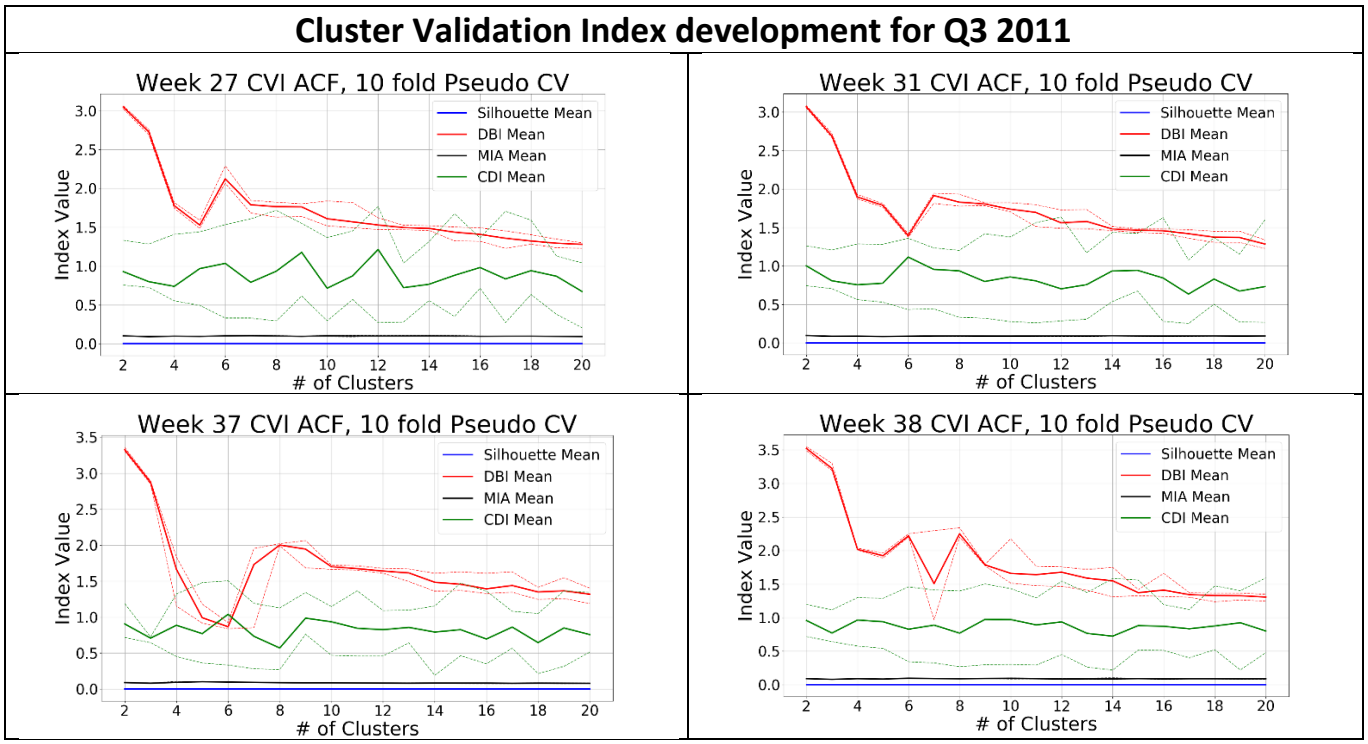


Figure 6. CVI development for four different weeks in Q3. Only the DBI index is able to produce convincing arguments for optimum cluster selection. The four weeks are selected based on their distinct estimate of optimum clusters.

Q3 2011	
Week	**CVI Estimated Clusters (indicates alternative clusters)**
27	5
28	3 (8)
29	8
30	5
31	6 (4)
32	6 (4)
33	7 (4)
34	6 (4)

35	6 (4)
36	6 (4)
37	6
38	4

Table 9. CVI-estimated optimum number of clusters for each week in Q3. This quarter exhibits fluctuation in the estimated optimum clusters across the entire period. Many of the optimum estimates indicate alternative number of clusters, shown by ().

Week 27 Varatio overlap with rest of weeks in Q3

Week	27 to 27	27 to 28	27 to 29	27 to 30	27 to 31	27 to 32	27 to 33	27 to 34	27 to 35	27 to 36	27 to 37	27 to 38
Cluster 0	100%	16%	14%	19%	16%	17%	18%	16%	16%	15%	15%	15%
Cluster 1	100%	10%	10%	11%	6%	8%	7%	7%	7%	7%	8%	7%
Cluster 2	100%	4%	3%	3%	2%	2%	1%	3%	2%	4%	5%	5%
Cluster 3	100%	7%	6%	7%	4%	6%	6%	5%	5%	5%	6%	6%
Cluster 4	100%	7%	9%	7%	7%	8%	9%	9%	8%	9%	9%	8%
Cluster 5	100%	9%	12%	11%	9%	8%	9%	8%	9%	8%	8%	9%

Week 32 Varatio overlap with rest of weeks in Q3

Week	32 to 27	32 to 28	32 to 29	32 to 30	32 to 31	32 to 32	32 to 33	32 to 34	32 to 35	32 to 36	32 to 37	32 to 38
Cluster 0	12%	10%	10%	12%	7%	100%	9%	8%	8%	8%	9%	9%
Cluster 1	10%	7%	9%	10%	8%	100%	9%	8%	9%	8%	9%	9%
Cluster 2	9%	8%	7%	9%	5%	100%	6%	6%	6%	6%	7%	6%
Cluster 3	5%	3%	2%	4%	3%	100%	5%	4%	4%	3%	6%	5%
Cluster 4	11%	8%	9%	9%	10%	100%	13%	13%	11%	12%	11%	11%
Cluster 5	20%	15%	14%	21%	17%	100%	21%	19%	18%	17%	16%	18%

Week 37 Varatio overlap with rest of weeks in Q3

Week	37 to 27	37 to 28	37 to 29	37 to 30	37 to 31	37 to 32	37 to 33	37 to 34	37 to 35	37 to 36	37 to 37	37 to 38
Cluster 0	12%	10%	10%	11%	7%	9%	9%	8%	8%	8%	100%	9%
Cluster 1	10%	7%	8%	9%	7%	7%	9%	8%	9%	8%	100%	10%
Cluster 2	7%	5%	2%	6%	3%	5%	4%	6%	3%	5%	100%	4%
Cluster 3	9%	7%	7%	6%	6%	8%	10%	10%	10%	12%	100%	11%
Cluster 4	8%	8%	7%	8%	5%	7%	6%	6%	7%	6%	100%	7%
Cluster 5	20%	15%	14%	20%	18%	19%	21%	20%	21%	21%	100%	21%

Table 10. Varatio for each cluster combination in weeks 27, 32, and 37. Dark green indicates a 50%+ Varatio coefficient. Light green indicates Varatio estimated at between 20-50%, yellow at 10-20%, light red at 5-10% and dark red at <5% of maximum variance as defined by Varatio. In all three selected weeks of Q3 the mapping is approximately 1:n.

5.4.Stability of Clusters in Quarter 4

The optimum number of clusters estimated across the thirteen weeks included from quarter 4 ranges from four to six. Figure 7 presents four different weeks of Q4 selected because of their distinct CVI development, while Table 11 shows each week's optimum cluster estimate. The selected number of cluster optima across Q4 is five and six, with six selected for easy comparison with Q1, Q2 and Q3. Table 12 shows the Varatio coefficients for three weeks in Q4, indicating a 1:n mapping of weekly clusters in Q4.

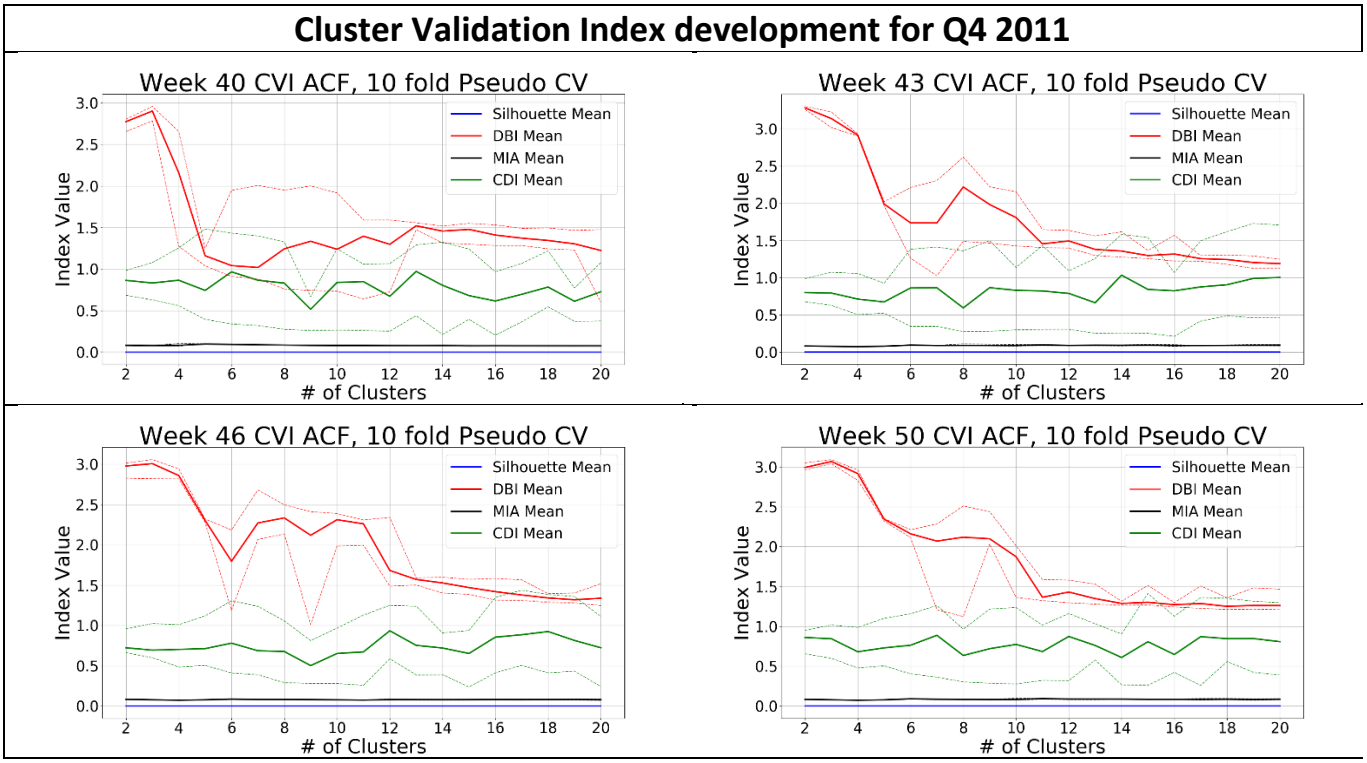


Figure 7. CVI development for four different weeks in Q4. Only the DBI index is able to produce convincing arguments for optimum cluster selection. The four weeks are selected based on their different estimates of optimum clusters and illustrate the variability of the 4th quarter.

Q4 2011	
Week	**CVI Estimated Clusters (indicates alternative clusters)**
39	4
40	5
41	6
42	6
43	6
44	5
45	6
46	6
47	6

48	5
49	5
50	5
51	5

Table 11. CVI-estimated optimum number of clusters for each week in Q4. This quarter has more stable defined estimates of optimum clusters across the entire period. Cluster solutions with five and six are equally prevalent, whereas in Q1, Q2 and Q3 the solution suggested six clusters.

Week 40 overlap with rest of weeks in Q4

Week	40 to 39	40 to 40	40 to 41	40 to 42	40 to 43	40 to 44	40 to 45	40 to 46	40 to 47	40 to 48	40 to 49	40 to 50	40 to 51
Cluster 0	23%	100%	21%	16%	19%	15%	14%	14%	13%	13%	12%	12%	14%
Cluster 1	11%	100%	9%	7%	8%	7%	7%	6%	7%	7%	7%	7%	10%
Cluster 2	10%	100%	10%	7%	9%	10%	9%	10%	9%	9%	9%	9%	9%
Cluster 3	6%	100%	5%	4%	4%	2%	4%	3%	3%	4%	5%	2%	5%
Cluster 4	10%	100%	12%	11%	13%	17%	14%	16%	15%	15%	16%	14%	10%
Cluster 5	7%	100%	6%	6%	6%	6%	6%	6%	6%	7%	7%	6%	9%

Week 46 overlap with rest of weeks in Q4

Week	46 to 39	46 to 40	46 to 41	46 to 42	46 to 43	46 to 44	46 to 45	46 to 46	46 to 47	46 to 48	46 to 49	46 to 50	46 to 51
Cluster 0	7%	10%	9%	9%	11%	15%	15%	100%	16%	15%	15%	15%	10%
Cluster 1	11%	9%	9%	8%	9%	8%	8%	100%	8%	9%	8%	8%	11%
Cluster 2	9%	9%	9%	7%	9%	11%	10%	100%	10%	11%	10%	10%	9%
Cluster 3	22%	19%	20%	18%	21%	18%	17%	100%	18%	17%	16%	15%	17%
Cluster 4	5%	5%	4%	2%	5%	2%	5%	100%	4%	5%	5%	3%	3%
Cluster 5	8%	7%	6%	6%	7%	7%	7%	100%	7%	8%	7%	7%	9%

Week 50 overlap with rest of weeks in Q4

Week	50 to 39	50 to 40	50 to 41	50 to 42	50 to 43	50 to 44	50 to 45	50 to 46	50 to 47	50 to 48	50 to 49	50 to 50	50 to 51
Cluster 0	23%	21%	21%	18%	21%	18%	17%	19%	19%	19%	19%	100%	20%
Cluster 1	6%	8%	8%	8%	9%	13%	13%	13%	14%	14%	17%	100%	11%
Cluster 2	9%	8%	8%	7%	8%	10%	9%	9%	10%	11%	11%	100%	10%
Cluster 3	7%	7%	6%	6%	6%	6%	7%	6%	7%	8%	7%	100%	9%
Cluster 4	3%	5%	6%	4%	4%	3%	3%	2%	3%	4%	4%	100%	5%
Cluster 5	10%	9%	9%	8%	8%	7%	8%	7%	8%	9%	9%	100%	12%

Table 12. Varatio for each cluster combination with week 40. Dark green indicates a 50%+ Varatio coefficient. Light green indicates Varatio estimated at between 20-50%, yellow at 10-20%, light red at 5-10% and dark red at <5% of maximum variance as defined by Varatio. In all three selected weeks of Q4 the mapping is 1:n.

5.5. Summarizing the Results

The analysis of consumption cluster stability and subsequent generalizability shows that weekly consumption clusters produce unstable clusters regardless of the season. Varatio shows that the mapping from one cluster to the remaining weeks in each quarter is approximately 1:n mapping. This indicates that the clustering is highly influenced by the specific week being clustered and that the clusters created for one week are not

applicable to any other week in that quarter. Though weeks were only compared within the same season, there is no reason to believe that the results would differ by comparing all weeks across the year.

6. Discussion

This paper has developed a novel method, Varatio, for evaluating the generalizability of consumption clusters created from smart-meter data. Varatio uses variance as a means of reducing the number of comparison matrices required. The method can produce a ratio of expected variance to observed variance. However, it is not able to produce better clusters or aid in the selection of clusters or algorithms, and thus is a tool applicable after the clusters have been selected using cluster validation indices. The aim of Varatio is to reduce the comparison matrix of two clusters into one vector encompassing the overlap information. This enables Varatio to generate a vector per cluster comparison and thus a matrix when comparing one week with all the remaining weeks rather than a matrix for comparison of two weeks.

Even though this paper has investigated cluster stability over the course of a year on a weekly basis and shows non-generalizable clustering results, the decision to choose weekly clustering can be contested. It is entirely possible that weekly clustering is not the optimum path for analyses of electricity consumption, as there may be too much variation imbedded in households' weekly behavior. Separating weekday and weekend consumption is an interesting prospect and may show that these are two very distinct entities which should be analyzed separately.

This investigation of the applicability of weekly clustering can easily be extended to contain different time intervals, or an even more complex division of recordings. This paper makes no assumption that the internal findings are generalizable to different time periods. The results show that, even though K-Means readily creates consumption clusters from smart-meter electricity data, the resulting clusters cannot be generalized, indicating that further reflection is needed for purposes of choosing which period should be used for clustering.

This paper includes and analyzes consumption data from more than 26,000 households for an entire year, but with only one recording per meter per hour in every week. The inclusion of data from consecutive years would allow this analysis to be performed across years such that any week could be reanalyzed, allowing for estimates of household variation between years. When it comes to producing knowledge about the stability of individual household consumption patterns, the Varatio tool is able to evaluate this type of clustering as well.

Furthermore, it is entirely conceivable that applications of other clustering algorithms can create clustering solutions which are generalizable. K-Means was selected because of its widespread application in smart-meter electricity-consumption clustering. The review in [4] discusses the large within-cluster variance resulting in overlapping clusters which fail to be statistically distinct. Without controlling the internal cluster variance in the clustering algorithm, the consumption clusters thus created will continue to be indistinguishable and impossible to generalize.

Although Varatio was developed for evaluating the generalizability of clusters, it does not aid in the selection of the appropriate clustering algorithm. Paper [10] showed that smart-meter data contains intrinsic structures, information that must be taken into account in improving the applicability of consumption clustering.

Finally, adjusting the recording frequency might also influence the outcome of clustering. This paper used sixty-minute recording intervals. Some data sets might contain different recording windows, which could potentially influence the generalizability of the clustering solution.

The smart-meter electricity consumption data are recorded for billing purposes but are successfully applied in many research papers for consumption clustering. This paper has shown that the ability of K-Means to create consumption clusters from smart-meter data does not convert into clustering solutions that are only valid in the data context that created them. The perceived ability of K-Means to cluster smart-meter consumption data does not translate into generally applicable clusters. Utilities cannot expect clusters created by employing K-Means to produce stable consumption patterns applicable in a business case setting.

7. Conclusion

A novel method for analyzing cluster stability has been developed, enabling this paper to show that clusters created by applying K-Means to smart-meter consumption data are not stable between weeks.

The Varatio method has been developed and applied as a tool for evaluating the stability of smart-meter consumption clusters at a weekly resolution for the whole of 2011. The results show that clusters from one week are not mapped 1:1 on to any other week within the same quarter. This finding is important in evaluating the generalizability and applicability of clusters created from a randomly selected week. The clusters are at best 20% identical across the weeks in any quarter, meaning that 80% of meters within a cluster rapidly disperse to other clusters in subsequent weeks.

Not only does the random initialization of K-Means induce the probability that the clustering is suboptimal – the difference in consumption between weeks, even within a quarter of a year, suggests that the clustering is highly dependent on initial decisions about which week to cluster. Varatio indicates that the consumption clusters created using K-Means are only stable for the week in which they are created. Clustering of smart-meter electricity data via K-Means suggests that clustering is academically achievable, but the generalizability of the clusters and their consequent practical applicability are another matter.

In the context of Danish electricity consumption clustering, this means that the preferred method is not able to create practically applicable clusters. The random initialization of the K-Means method induces problems which, through repeated initialization, are likely to result in good clustering, though the randomness between weeks is not managed at all.

Even though [10], [11] discuss ways of enabling K-Means to cluster time-series data, improvements to K-Means are needed if it is to produce viable cluster solutions from such types of data. Alternatively, methods of time-series clustering must be developed and applied when creating consumption clusters from smart-meter

consumption data. It is debatable whether the consumption clusters created from K-Means are able to extract consumption structures.

Supplementary Materials: The smart-meter electricity data analyzed in this paper is deemed sensitive and cannot be disclosed.

Acknowledgments: This work is part of the CITIES project funded in part by the Danish Innovation found. Grant DSF 1305-00027B. The data was provided by SydEnergi, with a special thanks to Fannar Thordarson at Ørsted for facilitating the contact to SydEnergi and Emil Mahler Larsen at Dansk Energi for rerunning the data extraction.

Conflicts of Interest: The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

[1] J. Worland, “Your Utility Company Wants to Sell You More than Just Electricity,” *TIME*, 2016. [Online]. Available: http://time.com/4312285/utility-company-electricity-solar-power/. [Accessed: 13-Jul-2018].

[2] S. Z. Christophe Guille, “How Utilities Are Deploying Data Analytics Now,” *Bain & Company*, 2016. [Online]. Available: http://www.bain.com/publications/articles/how-utilities-are-deploying-data-analytics-now.aspx. [Accessed: 13-Jul-2018].

[3] J. Mazurek, “The data treasure chest: Is there a market to sell utility data?,” *Accenture Blog*, 2016. [Online]. Available: https://www.accenture.com/us-en/blogs/blogs-utility-data-treasure-chest-there-market-sell-utility-data. [Accessed: 13-Jul-2018].

[4] M. Tureczek. Alexander and S. Nielsen. Per, “Structured Literature Review of Electricity Consumption Classification Using Smart Meter Data,” *Energies*, vol. 10, no. 5, p. 584, 2017.

[5] J. Kang and J. Lee, “Electricity Customer Clustering Following Experts’ Principle for Demand Response Applications,” *Energies*, vol. 8, pp. 12242–12265, 2015.

[6] J. D. Rhodes, W. J. Cole, C. R. Upshaw, T. F. Edgar, and M. E. Webber, “Clustering analysis of residential electricity demand profiles,” *Appl. Energy*, vol. 135, pp. 461–471, 2014.

[7] SEAS-NVE, “SEAS-NVE Watts.” [Online]. Available: https://watts.seas-nve.dk/?lang=en. [Accessed: 18-Jun-2018].

[8] S. Haben and C. Singleton, “Analysis and Clustering of Residential Customers Energy Behavioral Demand Using Smart Meter Data,” pp. 1–19, 2015.

[9] J. P. Gouveia and J. Seixas, “Unraveling electricity consumption profiles in households through clusters : Combining smart meters and door-to-door surveys,” *Energy Build.*, vol. 116, pp. 666–676, 2016.

[10] A. Tureczek, P. S. Nielsen, and H. Madsen, “Electricity Consumption Clustering Using Smart Meter Data,” *Energies*, vol. 11, no. 4, p. 859, 18AD.

[11] A. Tureczek, “Clustering District Heat Exchange Stations Using Smart Meter Consumption Data,” in *3rd International Conference on Smart Meter Energy Systems and 4th Generation District Heating*, 2017, p. 24.

[12] C. Beckel and T. Staake, “Are domestic load profiles stable over time ? An attempt to identify target households for demand side management campaigns,” pp. 4733–4738.

[13] D. Fjernvarme, “Fakta om fjernvarme,” 2017. [Online]. Available: http://www.danskfjernvarme.dk/presse/fakta-om-fjernvarme. [Accessed: 16-Mar-2018].

[14] S. Park, S. Ryu, Y. Choi, J. Kim, and H. Kim, “Data-Driven Baseline Estimation of Residential Buildings for Demand Response,” *Energies*, vol. 8, pp. 10239–10259, 2015.

[15] A. Ozawa, R. Furusato, and Y. Yoshida, “Determining the relationship between a household ’ s lifestyle and its electricity consumption in Japan by analyzing measured electric load profiles,” *Energy Build.*, vol. 119, pp. 200–210, 2016.

[16] S. Ramos, J. M. Duarte, F. J. Duarte, and Z. Vale, “A data-mining-based methodology to support MV electricity customers ’ characterization,” *Energy Build.*, vol. 91, pp. 16–25, 2015.

[17] R. Granell, C. J. Axon, and D. C. H. Wallom, “Impacts of Raw Data Temporal Resolution Using Selected Clustering Methods on Residential Electricity Load Profiles,” *IEEE Trans. Power Syst.*, vol. 30, no. 6, pp. 3217–3224, 2015.

[18] R. Al-otaibi, N. Jin, T. Wilcox, and P. Flach, “Feature Construction and Calibration for Clustering Daily Load Curves from Smart-Meter Data,” *IEEE Trans. Ind. Informatics*, vol. 12, no. 2, pp. 645–654, 2016.

[19] P. O. Perry, “Cross-Validation for Unsupervised Learning,” no. September, 2009.

[20] J. L. Viegas, S. M. Vieira, R. Melício, V. M. F. Mendes, and J. M. C. Sousa, “Classification of new electricity customers based on surveys and smart metering data Commission for Energy Regulation,” *Energy*, vol. 107, pp. 804–817, 2016.

[21] K. Basu, V. Debusschere, A. Douzal-chouakria, and S. Bacha, “Time series distance-based methods for non-intrusive load monitoring in residential buildings,” *Energy Build.*, vol. 96, pp. 109–117, 2015.

[22] J. Kwac, J. Flora, and R. Rajagopal, “Household energy consumption segmentation using hourly data,” *IEEE Trans. Smart Grid*, vol. 5, no. 1, pp. 420–430, 2014.

[23] F. M. Andersen, H. V Larsen, and T. K. Boomsma, “Long-term forecasting of hourly electricity load : Identification of consumption profiles and segmentation of customers,” *Energy Convers. Manag.*, vol. 68, pp. 244–252, 2013.

[24] G. Chicco, R. Napoli, and F. Piglione, “Comparisons Among Clustering Techniques for Electricity Customer Classification,” *IEEE Trans. Power Syst.*, vol. 21, no. 2, pp. 1–7, 2006.

[25] G. Chicco and J. Sumaili Akilimali, “Renyi entropy-based classification of daily electrical load patterns,” *IET Gener. Transm. Distrib.*, vol. 4, no. 6, pp. 736–745, 2010.

[26] J. J. López, J. A. Aguado, F. Martín, F. Mu, A. Rodríguez, and J. E. Ruiz, “Hopfield – K -Means clustering algorithm : A proposal for the segmentation of electricity customers,” vol. 81, pp. 716–724, 2011.